UNIVERSIDAD NACIONAL DE INGENIERÍA
RECINTO UNIVERSITARIO "SIMÓN BOLÍVAR"
FACULTAD DE ELECTROTECNIA Y COMPUTACIÓN

TRABAJO MONOGRÁFICO

"Implementación de Big Data en la información tributaria de la Dirección General de Ingresos en el área de Fiscalización"

PARA OPTAR AL TÍTULO DE
INGENIERO EN COMPUTACIÓN

ELABORADO POR:
Br. Josseling Jasmina Alemán Guido

TUTOR:
Ing. Gabriel Lacayo Saballos

MANAGUA, NICARAGUA
OCTUBRE 2017

DEDICATORIA

A Dios, por guiarme, ser mi fortaleza y cuidarme en los momentos más difíciles de este arduo caminar.

A mi Madre, por ser mi mayor apoyo y demostrarme todo su amor desde el momento en que nací hasta la fecha, por confiar en mi capacidad intelectual y saber que cada logro que hago es por brindarle un mejor futuro, por todos sus consejos y oraciones para convertirme en la mujer que soy y siempre ser el orgullo de ella.

A mi Padre, por ser el pilar del hogar y nunca dejarme desamparada.

A mis Hermanos por todo su apoyo durante los momentos que lo ameritaban, en especial a mi hermana Selena Alemán por tantos momentos compartidos y sus oraciones diarias por mí.

A mis Familiares más cercanos, que en algún momento me brindaron su apoyo para culminar mi carrera y siempre me brindaron un consejo en el momento más adecuado.

Al matrimonio Lacayo-Navarro por sus muestras de afectos y motivaciones para lograr ser una mejor persona cada día.

A la Institución Dirección General de Ingresos por abrir las puertas a mi crecimiento como profesional y desarrollarme de forma íntegra en el ámbito laboral.

Y a Todas las personas que a lo largo de este camino y hasta la fecha han sido un apoyo directo para mi persona y que con su amistad, basta para sacarme sonrisas en los momentos más difíciles, a todos ellos muchas gracias.

Josseling Alemán.

RESUMEN

La gran cantidad de datos almacenados y la necesidad de manipular la información de forma ágil para la toma de decisiones, es considerado como uno de los objetivos estratégicos que persigue la Dirección General de Ingresos (DGI) la cual se encarga de la recaudación de los tributos que permiten el desarrollo de nuestro país.

En el pasado se decía que la información es poder, pero en la actualidad el cómo se usa esta información es lo que hace la diferencia. En este documento se pretende que el lector conozca el concepto de Big Data y los conceptos tecnológicos que lo rodean, desde el punto de vista de software y hardware que puede llegar a variar dependiendo de la necesidad.

Se propuso un análisis y diseño para la implementación de una solución Big Data para el área de Fiscalización de la DGI; la cual permita hacerle frente a los inconvenientes concurrentes de análisis de información de los contribuyentes; también se presenta, las bases de datos no relacionales: NoSQL; se toma en consideración el uso de las tecnologías MongoDB como herramienta integrada al sistema de base de datos, y el framework Pentaho cuya finalidad será el análisis de reportes establecidos según el área de trabajo.

Esta solución de Big Data presentada al Centro Nacional de Datos Fiscales (CNDF), tiene como beneficios directos para la DGI:

1. Reducción óptima de reportes tabulares para obtener distintas vistas de información en el análisis de declaraciones tributarias

2. Facilidad en los procesos de selección, consulta y obtención de información personal para los contribuyentes.

3. Capacidad de analizar información dinámica e interactivamente.

4. Mejoría en los tiempos de respuestas del sistema para obtención de información.

INDICE

Índice de Ilustraciones

Índice de Tablas

1. INTRODUCCIÓN

En la actualidad el análisis de datos en las empresas es una necesidad cotidiana para guiar la toma de decisiones, obtener resultados centrados en el cliente, aprovechar los datos internos y crear un mejor ecosistema de información, es por estas razones, que se han venido creando e innovando innumerables herramientas donde el objetivo fundamental es dotar de una infraestructura tecnológica a las empresas y organizaciones, con la finalidad de gestionar la explosión de datos y de esta forma obtener información útil y eficiente para brindar mejores resultados económicos y operativos.

Una de estas innumerables herramientas para procesar grandes volúmenes de datos es “Big Data”, la cual ofrece la oportunidad de redefinir e inventar nuevos modelos de negocio, así como nuevos productos y servicios, además de obtener un mejor manejo de grandes cantidades de información lo que permite administrar y gestionar búsqueda de patrones concurrentes para la toma de decisiones.

El presente proyecto tiene como finalidad implementar una solución de Big Data, la cual apoye los esfuerzos de la Dirección General de Ingresos (DGI) en el análisis de datos para mejorar los procesos en cuanto a la reducción de riesgos y pérdidas en la información tributaria en el área de fiscalización, la cual afecte directamente en los procesos operativos de dicha institución.

En las restantes secciones de este documento se presenta la información que define el alcance y las pautas consideradas para desarrollar la solución propuesta. Se organiza en: Justificación, Objetivos, Marco teórico, Diseño Metodológico y Bibliografía, según se establece en la normativa vigente de la Universidad.

2. OBJETIVOS

2.1. Objetivo General

Implementar una solución basada en Big Data en el Centro Nacional de Datos Fiscales la cual facilite el análisis, seguimiento y control en la información tributaria del área de Fiscalización en la Dirección General de Ingresos.

2.2. Objetivo Específicos

1. Identificar las tecnologías, herramientas de software y requerimientos de hardware necesarios para la implementación de un ambiente de Big Data.

2. Seleccionar la técnica de modelado más adaptable para la configuración e integración de las herramientas en la construcción del ambiente de Big Data.

3. Realizar pruebas en la implementación sobre la base del diseño elaborado, aplicando los métodos y técnicas de la ingeniería de software que permitan el aseguramiento de la calidad del producto desarrollado.

4. Poner en producción la solución de Big Data y evaluar los resultados para verificar el cumplimiento de objetivos y beneficios esperados.

3. JUSTIFICACIÓN

Todas las organizaciones enfrentan el desafío de construir una arquitectura integral de información sobre riesgos, para superar estos desafíos las empresas deben adoptar un enfoque que les permita construir tanto confianza como valor entre la incertidumbre, para tener la capacidad de tomar decisiones viables y lograr un desempeño ajustado al riesgo para el negocio, cumpliendo en todo momento con estrictos requisitos regulatorios.

Las tecnologías de Big Data pueden reducir costos y aumentar los márgenes operativos mejorando la eficiencia de procesos, sobre una combinación optimizada de modelos predictivos, nuevas fuentes de datos y reglas estratégicas de negocio. Al optimizar sus operaciones, las organizaciones también pueden estar mejor preparadas para identificar e investigar acciones anómalas.

Actualmente, en el área de Fiscalización no existe ninguna herramienta informática que ayude a resolver la problemática en cuanto a: *el análisis de las declaraciones de los contribuyentes* y así de este modo evitar pérdidas a gran escala, o pequeñas reclamaciones en las declaraciones de los contribuyentes, por tanto las técnicas tradicionales tales como las acciones legales y la investigación privada requieren demasiado tiempo y dinero, además genera otro problema: una escasa eficiencia operativa. Hasta el momento solamente se llevan reportes tabulares con salida en formato documento Excel.

4. MARCO TEORICO

4.1. Definición del Big Data

Una de las aproximaciones más completas de Big Data es la facilitada por (GARTNER, 2013): “Son activos de información caracterizados por su alto volumen, velocidad y variedad, que demandan soluciones innovadores y eficientes para la mejora de conocimiento y toma de decisiones en las organizaciones”.

En términos generales podríamos referirnos a la tendencia en el avance de la tecnología, la cual es utilizada para describir enormes cantidades de datos de tal manera que, el concepto de Big Data aplica para toda aquella información que no puede ser procesada o analizada utilizando procesos o herramientas tradicionales. Sin embargo, Big Data no se refiere a alguna cantidad en específico, ya que es usualmente utilizado cuando se habla en términos de petabytes y exabytes de datos.

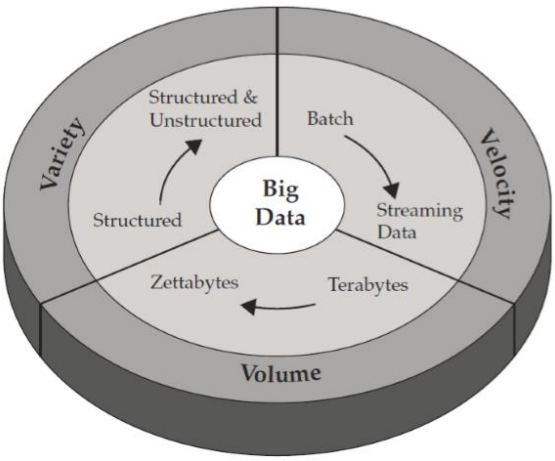

Ilustración 1. Definición de Big Data

No obstante, no hay que olvidarse de los inconvenientes del Big Data. Siendo el principal de ellos el proceso de adopción de Big Data: software, hardware necesario y su coste. Pero además existen otros muchos según (García, 2012) de menos peso como:

- Rechazo por parte del personal.
- Gasto de formación.
- Coste
- Problemas de privacidad.
- Problemas de información desactualizada.

En Big Data existen 3 tipos de datos, según (Services, 2012):

- Datos estructurados: son aquellos datos que tienen bien definido su longitud y formato. Suelen ser fechas y números, cadenas de caracteres y están almacenados en tablas.
- Datos no estructurados: son lo opuesto a los datos estructurados, es decir, carecen de un formato específico. Estos datos estructurados son generados por: imágenes de satélites, datos científicos, fotografía y video, documentos, presentaciones, correos electrónicos, etc.
- Datos semi-estructurados: son una mezcla de los estructurados y no estructurados, es decir, estos datos siguen una especie de estructura implícita, pero no tan regular como para poder ser gestionada y automatizada como la información estructurada.

4.2. Importancia del Big Data

(Fragoso, 2012), considera que con el término Big Data se hace referencia a la tendencia del avance de las tecnologías que han abierto las puertas hacia un nuevo enfoque de entendimiento y toma de decisiones, por lo tanto, el Big Data se aplicará para toda aquella información que no pueda ser procesada por los métodos tradicionales.

Un sistema gestor de bases de datos (SGBD), según (Silberschatz, 2002), consiste en una colección de datos interrelacionados y un conjunto de programas para acceder a dichos datos. La colección de datos, normalmente denominada base de datos, contiene información relevante para una empresa. El objetivo principal de un SGBD es proporcionar una forma de almacenar y recuperar la información de una base de datos de manera que sea tanto práctica como eficiente. El gran inconveniente que presenta, es el tiempo necesario para manejar grandes cantidades de datos, pero esto se logra gracias al Big Data, debido a su estructura que es capaz de almacenar y procesar grandes volúmenes de datos, además es una arquitectura orientada a los programas actuales.

Una vez que hemos hecho referencia de la importancia de Big Data, sobre todo gracias a la mejora con respecto a los modelos relacionales se citarán los beneficios más habituales del Big Data (García, 2012), no obstante estos beneficios no se tienen porque aplicar a todas las organizaciones, ya que cada organización tiene y actúa en diferentes condiciones.

- Búsqueda de nuevas oportunidades de negocio a través de segmentación mejorada y venta cruzada de productos (mejora de la estrategia).

- Mejoras operativas: mayor capacidad de visibilidad del negocio a través de informes más detallados.

- Cuadro de mandos en tiempo real, la información siempre está disponible sin esperas de actualización de los datos (información en tiempo real).

- Permite la simplificación de procesos actuales y control de negocio (reducción de costes).

- Permite detectar patrones complejos de fraude en tiempo real analizando los datos históricos, análisis de transacciones y operaciones sospechosas (reducción de costes).

4.3. Características Big Data

Las principales características que reúne Big Data (García, 2012) y que destacan sobre todas las demás y que lo hacen ser único son:

- **Volumen:** suele utilizarse como sinónimo de Big Data, el reto relacionado con el volumen de datos se ha puesto de manifiesto recientemente, debido al crecimiento de los sistemas de información e inteligencia, el incremento del intercambio de datos entre sistemas y dispositivos nuevos, nuevas fuentes de datos y el nivel creciente de digitalización de los medios de comunicación que antes solo estaban disponibles en otros formatos, tales como texto, imágenes, videos y audio. Es por tal razón que, las empresas están cubiertas de una cantidad cada vez mayor de datos de todo tipo, acumulando fácilmente terabytes, incluso peta bytes, de información, (García, 2012).

- **Velocidad:** se asocia con la necesidad de utilizar los datos más rápidamente. Fuentes de datos automatizados, tales como sensores, GPS, etc, los cuales generan datos cada fracción de segundo, los dispositivos que generan datos a intervalos más largos, tales como los teléfonos inteligentes, también terminan generando corrientes constantes de datos que necesitan ser ingeridos rápidamente. Por otro lado, todos estos datos tienen poco o ningún valor si no se convierten en información útil, (García, 2012).
- **Variedad:** los grandes volúmenes de datos incluyen cualquier tipo de datos, estructurados y no estructurados como texto, datos de sensores, audio, video, secuencias de clic o archivos de registros, entre otros. Al analizar estos datos juntos se encuentra información nueva. Esta característica está relacionada con la organización de los datos, esta organización se divide básicamente en datos estructurados, semi-estructurados y no estructurados, (García, 2012).

A continuación se muestra de manera gráfica (Soubra, 2012), las tres características que reúne Big Data:

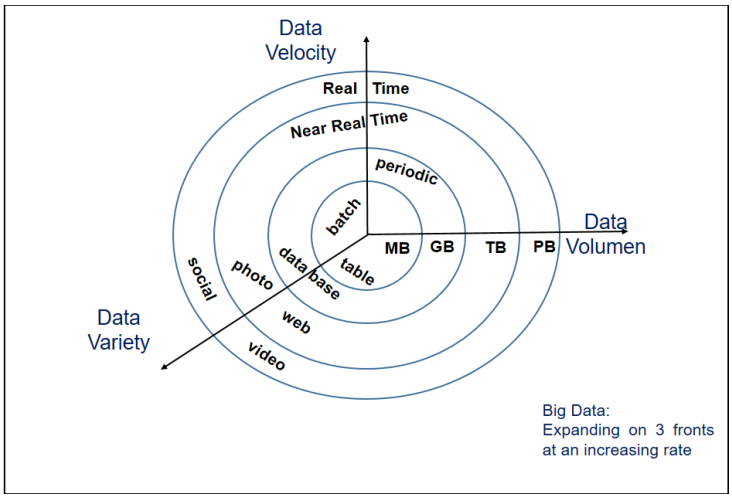

Ilustración 2. Las 3V del Big Data

4.4. Seguridad en Big Data

La seguridad importa tanto a las empresas debido la perdida de información que implica (pérdida de clientes, de información valiosa), también es de suma importancia para el consumidor que cada vez es más consciente de cómo es utilizada su información por ello exigen políticas de seguridad que en muchas ocasiones no se tienen en cuenta. Por todo esto es tan importante tener un buen sistema de seguridad para el Big Data (García, 2012).

Por lo tanto queda visible que es necesario intensificar la seguridad en los modelos de Big Data. Para ello ya existen modelos o softwares de seguridad específicos como, (García, 2012):

- Lookwise Solutions, está dedicada al desarrollo de productos desde hace 10 años, que dan respuesta las necesidades de las organizaciones en materia de gestión de la seguridad, Big Data y de cumplimiento normativo, (García, 2012).

- ISO / IEC 17799:2005, establece los lineamientos y principios generales para iniciar, implementar, mantener y mejorar la gestión de seguridad de la información en una organización, contiene las mejores prácticas de los objetivos de control, tales objetivos ofrecen orientaciones generales sobre las metas comúnmente aceptadas de gestión de seguridad de información, (García, 2012).

- COBIT, es precisamente un modelo para auditar la gestión y control de los sistemas de información y tecnología, orientado a todos los sectores

de una organización, es decir, administradores, usuarios y por supuesto, los auditores involucrados en el proceso, (García, 2012).

- ISO / IEC 27000, es un conjunto de estándares desarrollados que proporcionan un marco de gestión de la seguridad de la información utilizable por cualquier tipo de organización, pública o privada, grande o pequeña, (García, 2012).

4.5. Plataformas y software para tratamiento de Big Data

Para el manejo de datos en una solucion de Big Data según (García, 2012), es necesario tener dos componentes básicos, tanto el hardware como el software. En la actualidad existen diferentes herramientas para soluciones de Big Data (Foundation, Welcome to Apache Avro!, 2017), pero para hablar del software de tratamiento de grandes almacenes de datos nos vamos a centralizar en Bases de Datos NoSQL con la integración de Pentaho para el procesamiento y análisis de reportería.

4.5.1. Bases de Datos NoSQL

Cuando se habla de NoSQL (Acens), no nos referimos únicamente a un tipo de bases de datos sino a diferentes soluciones dadas para almacenar datos cuando las bases de datos relacionales nos generan problemas. Las bases de datos NoSQL son sistemas de almacenamiento de información que no cumplen con el esquema entidad-relación, no imponen una estructura de datos en forma de tablas y relaciones entre ellas , en ese sentido son más

flexibles, ya que suelen permitir almacenar información en otros formatos como: clave-valor, mapeo de columnas, documentos o grafos.

Además de lo comentado anteriormente, las bases de datos NoSQL son, (Acens) sistemas de almacenamiento de información que no cumplen con el esquema entidad – relación.

4.5.2. Teorema CAP

El Teorema CAP (Dev, s.f.), sostiene que es imposible que un sistema computacional distribuido pueda proveer las tres siguientes propiedades a la vez:

- **Consistencia:** al realizar una operación siempre se tiene que recibir la misma información, sin importar el nodo que procese el pedido. Significa que no importa el nodo que conforma nuestra base de datos reciba un pedido, todos deben responder ante la operación de igual manera y debe ser transparente para nosotros quién la efectuó. Todos los clientes ven la misma versión de los datos, (Dev, s.f.).
- **Disponibilidad:** el sistema garantiza respuestas para todos los requerimientos que recibe, aún si uno o más nodos se encuentran caídos, (Dev, s.f.).
- **Tolerancia a particiones:** el sistema sigue funcionado a pesar de que haya sido dividido por un fallo en la red, (Dev, s.f.).

Ilustración 3. Teorema CAP

4.5.2.1. Clasificación de cada base de datos NoSQL según el teorema CAP

Para ser escalables y distribuidas, las bases de datos NoSQL, siguen distintos métodos, por lo que no todas cumplen los mismos puntos del teorema CAP.

En la Ilustración 3, se observa cómo se desprenden cuatro espacios posibles en el cual se puede categorizar a cualquier motor de Base de Datos NoSQL (Dev, s.f.):

- **AP**: garantizan disponibilidad y tolerancia a particiones, pero no la consistencia, al menos de forma total. Algunas de ellas consiguen una consistencia parcial a través de la replicación y la verificación, (Dev, s.f.).

- **CP**: garantizan consistencia y tolerancia a particiones. Para lograr la consistencia y replicar los datos a través de los nodos, sacrifican la disponibilidad, (Dev, s.f.).
- **CA**: garantizan consistencia y disponibilidad, pero tienen problemas con la tolerancia a particiones. Este problema lo suelen gestionar replicando los datos, (Dev, s.f.).

Hay que tener en cuenta, que esta clasificación no es definitiva, ya que algunos de estos sistemas NoSQL pueden configurarse para cambiar su comportamiento. Por tanto, además de pensar en el tipo de base de datos NoSQL que mejor se adapta a nuestro modelo de datos, también tendremos que pensar en su funcionamiento; así podremos conseguir que nuestra aplicación funcione de la mejor manera posible, (Dev, s.f.).

4.5.3. Ventajas

Esta forma de almacenar la información ofrece ciertas ventajas sobre los modelos relacionales, entre las ventajas más significativas podemos destacar, (García, 2012):

Se ejecutan en máquinas con pocos recursos: estos sistemas, a diferencia de los sistemas basados en SQL, no requieren de apenas computación, por lo que se pueden montar en máquinas de un coste más reducido, (García, 2012).

Escalabilidad horizontal: para mejorar el rendimiento de estos sistemas simplemente se consigue añadiendo más nodos, con la única operación de indicar al sistema cuáles son los nodos que están disponibles, (García, 2012).

Pueden manejar gran cantidad de datos: Esto es debido a que utiliza una estructura distribuida, en muchos casos mediante tablas Hash, (García, 2012).

No genera cuellos de botella: El principal problema de los sistemas SQL es que necesitan transcribir cada sentencia para poder ser ejecutada, y cada sentencia compleja requiere además de un nivel de ejecución aún más complejo, lo que constituye un punto de entrada en común, que ante muchas peticiones puede ralentizar el sistema, (García, 2012).

4.5.4. Tipos de Bases de Datos NoSQL

Dependiendo de la forma en la que almacenen la información, se puede encontrar varios tipos de bases de datos NoSQL, (Acens). Se citan los más utilizados:

- ✓ Bases de Datos NoSQL Clave-Valor
- ✓ Bases de Datos NoSQL Documentales
- ✓ Bases de Datos NoSQL Grafo
- ✓ Bases de Datos NoSQL Columnares

A continuación se define cada una de estas:

4.5.4.1. Bases de Datos Clave-Valor

Son el modelo de base de datos NoSQL más popular, además de ser la más sencilla en cuanto a funcionalidad, (Acens). En este tipo de sistema, cada elemento está identificado por una llave única, lo que permite la recuperación de la información de forma muy rápida, información que habitualmente está almacenada como un objeto binario (BLOB). Se caracterizan por ser muy eficientes tanto para las lecturas como para las escrituras.

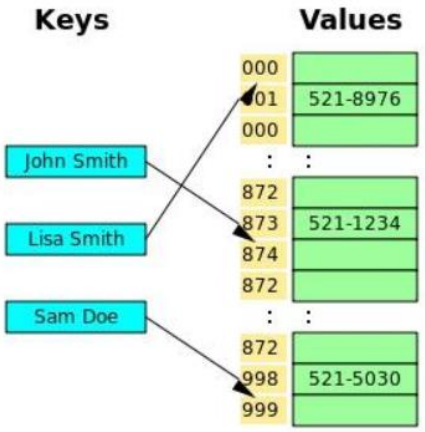

Ilustración 4. Bases de Datos Clave – Valor

4.5.4.2. Bases de Datos Documentales

Este tipo almacena la información como un documento, generalmente utilizando para ello una estructura simple como JSON o XML y donde se utiliza una clave única para cada registro, (Acens). Este tipo de implementación permite, además de realizar búsquedas por clave - valor, realizar consultas más avanzadas sobre el contenido del documento. Son las bases de datos NoSQL más versátiles.

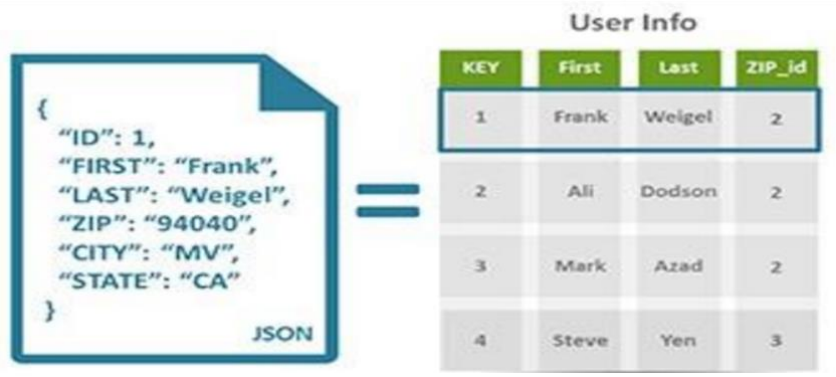

Ilustración 5. Bases de Datos Documentales

4.5.4.3. Bases de Datos en Grafo

En este tipo de bases de datos, la información se representa como nodos de un grafo y sus relaciones con las aristas del mismo, de manera que se puede hacer uso de la teoría de grafos para recorrerla, (Acens). Para sacar el máximo rendimiento a este tipo de bases de datos, su estructura debe estar totalmente normalizada, de forma que cada tabla tenga una sola columna y cada relación dos.

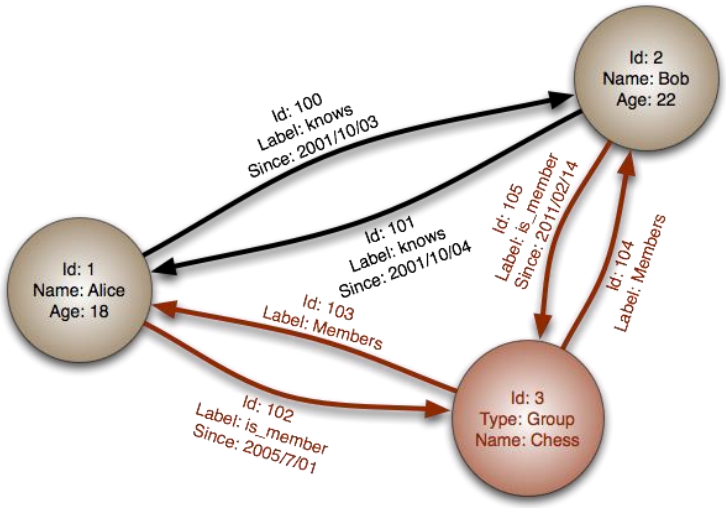

Ilustración 6. Bases de Datos en Grafo

4.5.4.4. Bases de Datos Columnares

Como su nombre lo indica, las bases de datos están organizados de columna por columna en lugar de la fila: es decir, todos los casos de un solo elemento de datos (por ejemplo, *nombre de cliente*) se almacenan de modo que se puede acceder como una unidad, (Garcete). Esto los hace especialmente eficaz en las consultas analíticas, como la lista de selecciones, que a menudo lee unos pocos elementos de datos, pero necesitamos ver todas las instancias de estos elementos.

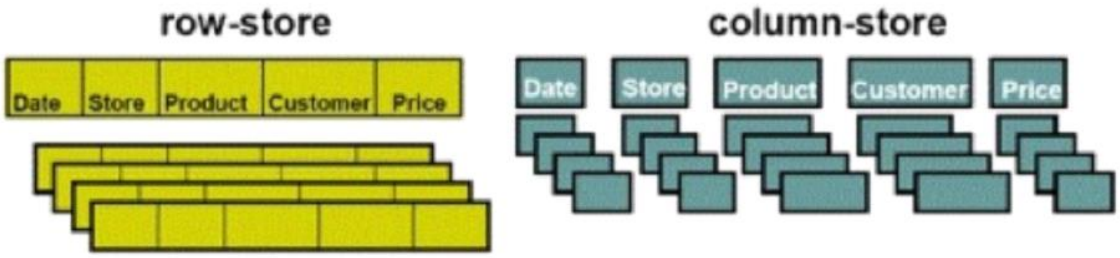

Ilustración 7. Bases de Datos Columnares

4.5.5. Pentaho

Pentaho es una herramienta de Business Intelligence desarrollada bajo la filosofía del software libre para la gestión y toma de decisiones empresariales. Es una plataforma compuesta de diferentes programas que satisfacen los requisitos de Business Intelligence. Ofreciendo soluciones para la gestión y análisis de la información, (Gravitar, s.f.).

La plataforma ha sido desarrollada bajo el lenguaje de programación Java y tiene un ambiente de implementación también basado en Java, haciendo así que Pentaho sea una solución muy flexible al cubrir una alta gama de necesidades empresariales.

Ilustración 8. Símbolo de Pentaho

4.5.5.1. Características de Pentaho

Con Pentaho podemos realizar, (Gravitar, s.f.):

- Procesos ETL (Extracción, Transformación y Carga de datos).
- Análisis de datos.
- Reportería para la empresa.
- Cuadros de mando para la toma de decisiones empresariales.
- Minería de datos (Data Mining) para el estudio y tratamiento de datos masivos.

4.5.5.2. Componentes de la Suite de Pentaho

4.5.5.2.1. PDI (Pentaho Data Integration)

Pentaho Data Integration abre, limpia e integra la información de manera consistente, en una sola versión de todos los recursos de información, que es uno de los más grandes desafíos para las organizaciones TI hoy en día. Pentaho Data Integration permite una poderosa ETL, (Gravitar, s.f.).

Kettle

Kettle(Gravitar, s.f.), es una herramienta de las que se denominan ETL (Extract – Transform – Load). Es decir, una herramienta de Extracción de datos de una fuente, Transformación de esos datos, y Carga de esos datos en otro sitio.

Estas tareas son típicas en procesos de migración, integración con terceros, *explotación de Big Data* y en general se podría decir que son necesarias en

casi cualquier proyecto mediano o grande. Por eso Kettle nace con la intención de facilitarnos este trabajo, de forma que no tengamos que entrar en el detalle de la implementación de como se hace cada una de estas tareas, sino que simplemente especificamos qué es lo que queremos hacer.

El uso de kettle permite evitar grandes cargas de trabajo manual frecuentemente difícil de mantener y de desplegar, (Gravitar, s.f.).

Ilustración 9. Símbolo de Kettle

Características

A parte de ser open source y sin costes de licencia, las características básicas de esta herramienta son, (Gravitar, s.f.):

- Entorno gráfico de desarrollo
- Uso de tecnologías estándar: Java, XML, JavaScript
- Fácil de instalar y configurar
- Multiplataforma: Windows, Macintosh, Linux
- Basado en dos tipos de objetos: Transformaciones (colección de pasos en un proceso ETL) y trabajos (colección de transformaciones)
- Incluye cuatro herramientas:
    - Spoon: para diseñar transformaciones ETTL usando el entorno gráfico.

- PAN: para ejecutar transformaciones diseñadas con spoon.
- CHEF: para crear trabajos.
- Kitchen: para ejecutar trabajos.

4.5.5.2.2. PRD (Pentaho Report Designer)

El Pentaho Report Designer es una herramienta que simplifica el proceso de generación de reportes, permitiendo a los diseñadores de reportes crear rápidamente informes sofisticados. El diseñador de reportes ofrece un entorno gráfico con herramientas intuitivas y fáciles de utilizar, y una estructura de reporte bastante acertada y flexible para darle libertad al diseñador de generar reportes que se adapten totalmente a su gusto y necesidad, (Gravitar, s.f.).

Características

A continuación se presentan las diferentes características que posee PRD, (LópMor, 2011):

- Diseñador gráfico basado en “arrastrar y soltar” (drag & drop) que provee completo control de acceso a los datos, agrupaciones, cálculos, gráficas, formato, etc. Para reportes de alta resolución.
- Asistente paso a paso integrado, que guía a los diseñadores de reportes durante el proceso de diseño.
- Plantillas de reportes, aceleran el proceso de generación, proporcionando un aspecto consistente y atractivo.
- Opciones de salida flexibles, incluyendo los populares formatos Adobe PDF, HTML, Microsoft Excel, entro otros.

5. ANALISIS Y PRESENTACION DE RESULTADOS

5.1. Metodología Empleada

Una parte sumamente importante para la implementación de la solución de Big Data, es el almacén de datos, que en este tipo de aplicaciones conforma el corazón del sistema. A fin de obtener los resultados que se plantean en los objetivos, nos apoyamos del proceso metodológico CRISP que se detallará a continuación, donde se estructura el ciclo de vida de un proyecto en seis fases, que interactúan entre ellas de forma iterativa durante el desarrollo del proyecto, (IBM, 2012).

El proceso de implementación que establece CRISP se describe en Ilustración 11, que se muestra a continuación:

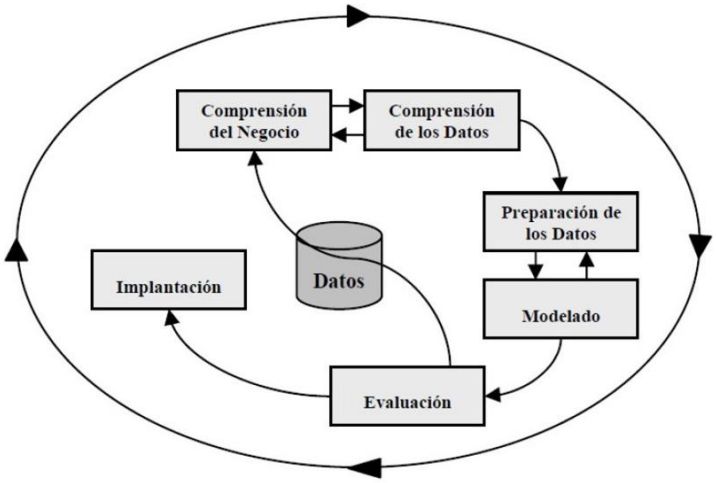

Ilustración 10. Metodología CRISP

5.2. Definición de roles del proyecto

En la Tabla 1, se describen los diferentes roles con sus respectivas responsabilidades, para la implementación de la solución de Big Data. Los roles asociados para la implementación del proyecto son

- Analista de Información
- Arquitecto de Información
- Cliente
- Usuario

Tabla 1. Descripción de Roles (Elaboración Propia)

Rol	Descripción	Responsabilidades
Analista de Información	Ing. José Herrera	1. Identificar problema y/o necesidad del cliente / usuario a partir de la elaboración de las preguntas de negocio. 2. Identificar fuentes de información internas y externas y constatar la veracidad de la información. 4. Identificar la periodicidad para la generación de resultados (informes).

		5. Identificar responsables por unidades de trabajo. 6. Velar por la usabilidad de las soluciones a implementar.
Arquitecto de Información	Ing. Josseling Alemán	1. Validación de accesos y permisos para las aplicaciones que intervengan dentro de la solución a implementar. 2. Identificar capacidades y opciones de almacenamiento. 3. Garantizar la seguridad y persistencia de la información. 4. Identificar las herramientas de software y hardware para el procesamiento de información. 5. Identificar herramientas para la generación de reportes y visualización grafica de la información.
Cliente	Dirección General de Ingresos	1. Apoyar el proceso de identificación de requisitos

		funcionales y no funcionales del negocio. 2. Identificar información necesaria y clave para el negocio. 3. Apoyar el proceso de validar la veracidad de la información.
Usuario	Personal de Área Fiscalización	1. Velar por la buena comunicación con el equipo de trabajo. 2. Apoyar procesos y tareas enfocadas en la usabilidad del proyecto una vez finalizado.

5.3. Desarrollo de la Solución

Según lo planteado anteriormente, el desarrollo e implantación de la solución se realizó siguiendo la metodología de CRISP, por lo cual se reporta en el presente documento de acuerdo al ciclo de vida de dicha metodología, según se detalla a continuación:

5.3.1. Primera Fase: Comprensión del Negocio

A continuación se define cada una de las tareas que componen esta fase:

5.3.1.1. Determinar los objetivos del negocio

La Dirección General de Ingresos, como parte integral de la administración tributaria, tiene el mandato de gestionar, controlar y recaudar el conjunto de los impuestos internos con equidad, transparencia y eficiencia, promoviendo la cultura tributaria y cumpliendo con el marco legal, aportando al gobierno recursos para el desarrollo económico y social del país. Al afrontar este desafío de avanzar hacia un mayor cumplimiento de las obligaciones impositivas, la Dirección General de Ingresos debe encontrar fórmulas para una mejora en las condiciones para la utilización productiva y social, así mismo incrementar el número de contribuyentes y el prestigio ante la sociedad, que es la destinataria ultima de los resultados de la gestión de la DGI. Una solución propuesta a esto es brindar un enfoque específico a la acción institucional del área de fiscalización de forma integral.

Además se ha encargado un estudio con los siguientes objetivos:

- Incrementar el número de contribuyentes potenciales para aumentar el porcentaje de recaudación anual.
- Detectar incumplimiento y evasión de pagos de impuestos.

5.3.1.2. Compilación de la información de la empresa

- Determinar la estructura de la organización

  A continuación se presenta la estructura organizacional de la Dirección General de Ingresos de tal forma que permita tener una visión de las áreas que están involucradas y que se verán beneficiadas con el proyecto.

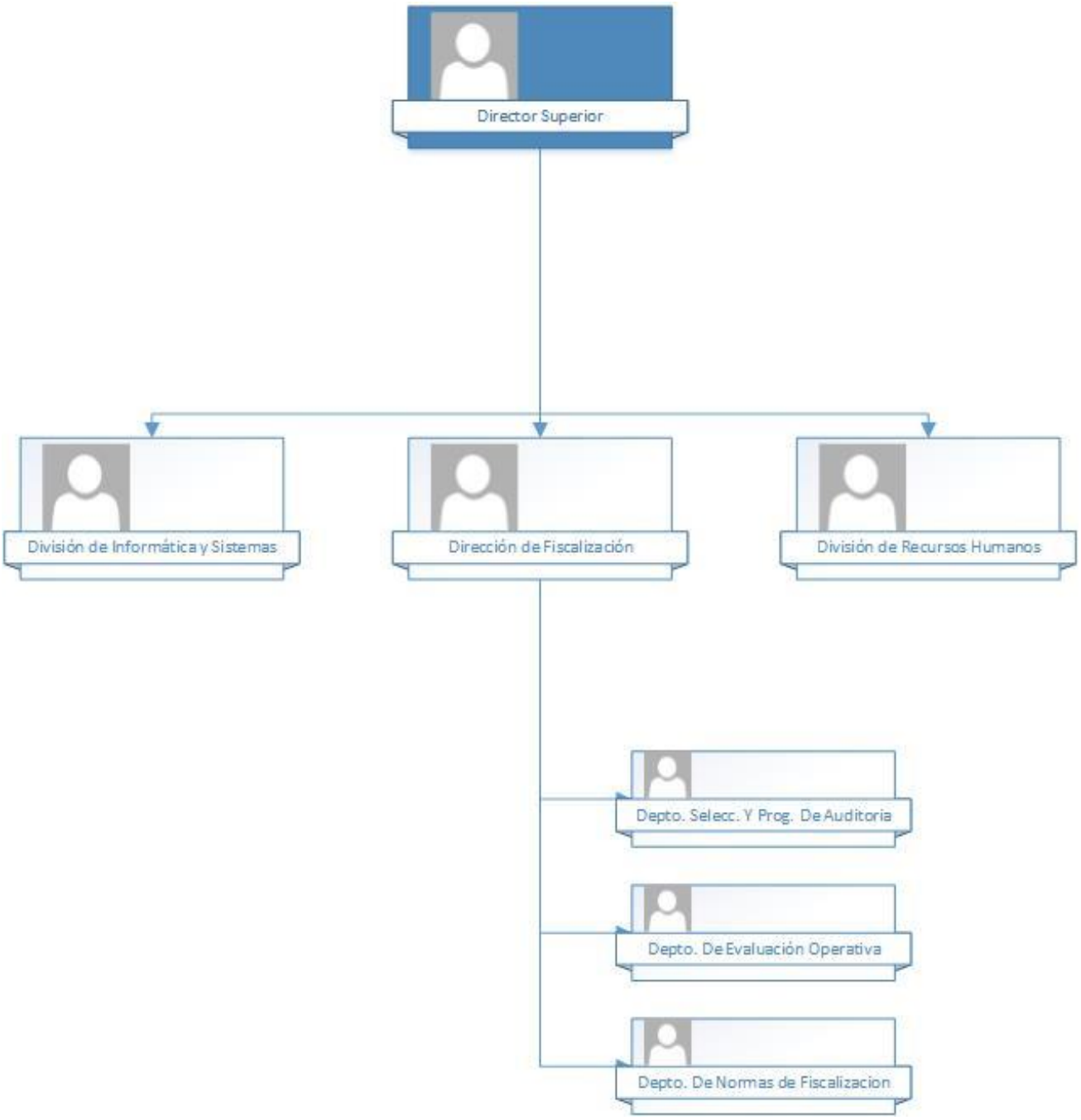

Ilustración 11. Organigrama DGI

5.3.1.3. Describir el área interesada

El área de la interesada a la que está orientada la solución de Big Data es: fiscalización, donde las actuales prácticas, recursos y diseño no le permiten alcanzar el modelo de control integral deseado.

Hasta ahora la Dirección General de Ingresos no ha podido dedicar recursos, para definir su estrategia de control global e identificar los riesgos de cumplimiento presente en cada una de las instancias de interacción con el contribuyente.

La función central de fiscalización de la Dirección General de Ingresos ha tenido que acometer simultáneamente tareas normativas y tareas de ejecución de auditorías, lo que lleva a una limitada dotación de fiscalizadores, lo cual no ayuda al cumplimiento del objetivo fundamental del área que es aplicar un modelo de fiscalización inteligente, que genere planes de fiscalización para los Grandes, Medianos y Pequeños Contribuyentes, permitiendo así un aumento significativo en la recaudación tributaria.

5.3.1.4. Describir la solución actual

En la actualidad no se dispone de una propuesta de solución ante la necesidad planteada.

5.3.1.5. Inventario de recursos

Actualmente en el Centro Nacional de Datos Fiscales, se cuenta con el equipo de hardware necesario para la implementación de la solución del proyecto de Big Data. El origen de datos a cargarse en el modelo de

implementación de Big Data, será a través de uno de los mayores sistemas actual implementado en la DGI: el sistema de Ventanilla Electrónica Tributaria VET.

En cuanto a recursos de software disponemos de un servidor virtual en Linux para la instalación de un Base de Datos MongoDB y un gestor de apoyo llamado RoboMongo, que es con la que contamos para el almacenamiento de los datos y la implementación de la solución de Big Data, en cuanto a reportería para el análisis de los datos se instaló Pentaho.

Tabla 2. Especificaciones Técnicas de Hardware (Elaboración Propia)

Máquina Virtual	Ubuntu 64-bit
Características	• 1 Procesador 2,30 GHz • 64 bits • 4.3GB RAM • 150GB disco duro • Ubuntu-14.04-server-amd64
Servicios Instalados	• Base de Datos NoSQL: MongoDB • Analisis de Información (Reporteria): Pentaho

La fuente de datos es una base de datos MySQL con la información de los contribuyentes inscritos en las Rentas de la Cuidad de Managua de la

Dirección General de Ingresos a nivel nacional en un período comprendido del año 2012 hasta el 2015.

5.3.2. Segunda Fase: Comprensión de los Datos

Se usará la información de los sistemas actuales, específicamente del sistema: VET (Ventanilla Electrónica Tributaria), que se encuentra alojado en un servidor de Base de Datos Relacional: MySQL. Existe gran cantidad de registros y atributos para procesar en la solución a implementarse con Big Data, aunque se haiga seleccionado cierta cantidad de datos de contribuyentes registrados para el estudio inicial de la solución a implementarse.

Se planificó una depuración de los datos, luego de ya haber obtenido la selección de los mismos, para solucionar cualquier problema al momento de la integración de los datos.

5.3.3. Tercera Fase: Preparación de los Datos

La preparación de los datos se hizo bajo ciertos criterios de selección de atributos: la base de datos (sairi) es uno de los sistemas integrados que hemos mencionado anteriormente; la cual contendrá información de contribuyentes, por lo que es importante filtrar los atributos como nombre, dirección, número de teléfono, numero RUC, renta, entre otros.

En la siguiente Ilustración 13, se observa el bosquejo de los datos organizados en un modelo relacional de base de datos en la plataforma MySQL.

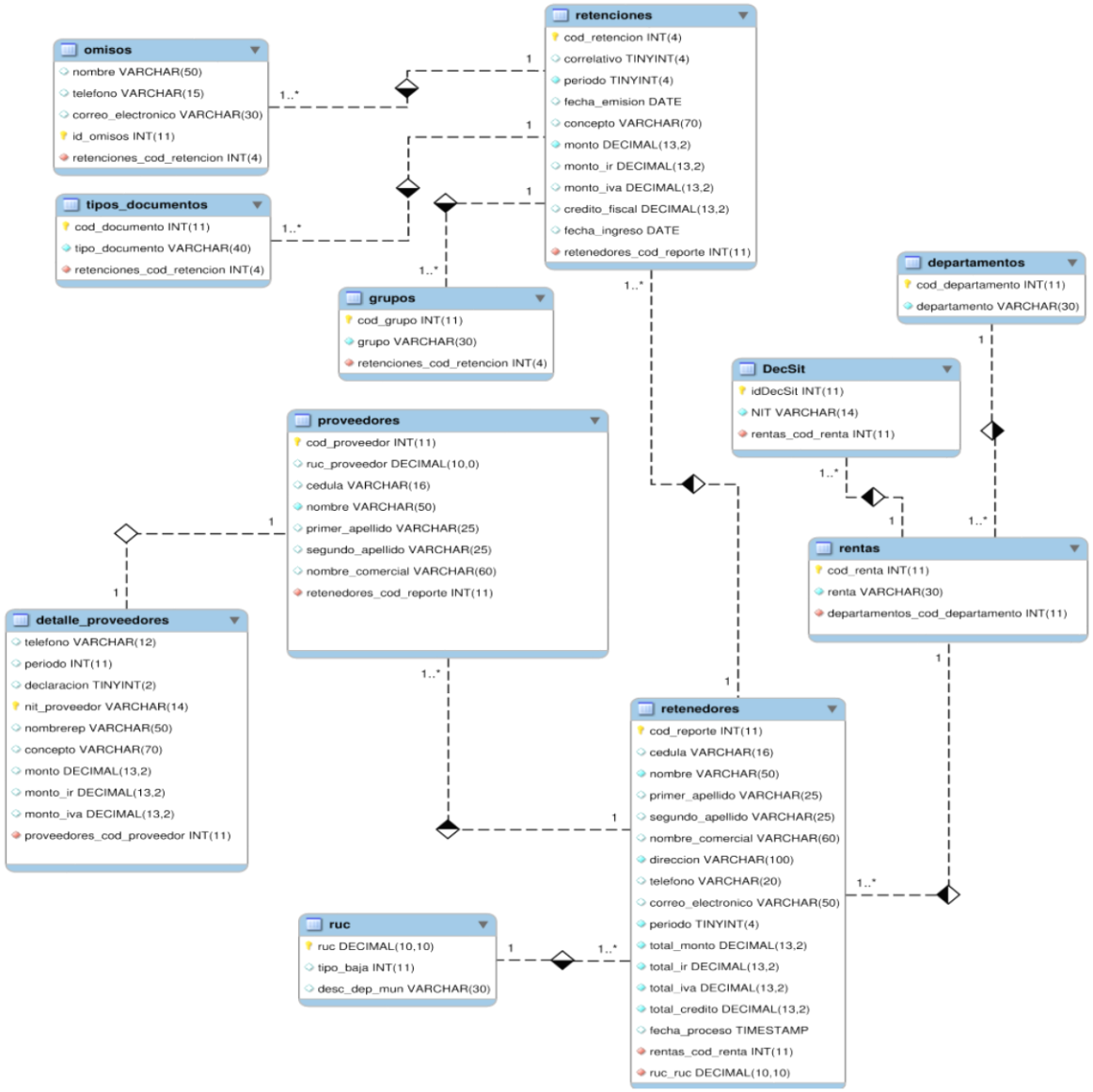

Ilustración 12. Modelo Relacional MySQL

A continuación describiremos cada una de ellas detallando cada uno de sus campos.

Tabla 3. DecSit

TABLA: DecSit		
Columna	**Tipo de Dato**	**Descripción**
idDecSit	Integer	Primary Key del código de contribuyente.
NIT	Varchar	Identificador de contribuyente
rentas_cod_renta	Integer	Foreing Key de la tabla Rentas

Tabla 4. Rentas

TABLA: rentas		
Columna	**Tipo de Dato**	**Descripción**
cod_renta	Integer	Primary Key del código de renta.
renta	Varchar	Nombre de la Renta al cual está adscrito el Contribuyente
departamentos_cod_departamento	Integer	Foreing Key de la tabla departamentos

Tabla 5. Departamentos

TABLA: departamentos		
Columna	**Tipo de Dato**	**Descripción**
cod_departamento	Integer	Primary Key del código de departamento.
Departamento	Varchar	Nombre del departamento al que pertenece el contribuyente.

Tabla 6. Retenedores

TABLA: retenedores		
Columna	**Tipo de Dato**	**Descripción**
cod_reporte	Integer	Primary Key del código de Reporte generado por el Contribuyente
cedula	Varchar	Número de Cedula del Contribuyente
nombre	Varchar	Nombre del Contribuyente
primer_apellido	Varchar	Primer Apellido del Contribuyente
segundo_apellido	Varchar	Segundo Apellido del Contribuyente
nombre_comercial	Varchar	Nombre Comercial del Negocio del Contribuyente
dirección	Varchar	Dirección del Negocio del contribuyente
teléfono	Varchar	Número de Teléfono del Contribuyente
correo_electronico	Varchar	Correo Electrónico del Contribuyente

periodo	Integer	Indica el Mes anexando el Año para el pago del periodo
total_monto	Decimal	Monto Total a pagar
total_ir	Decimal	Monto de IR a Declarar
total_iva	Decimal	Monto de IVA a Declarar
total_credito	Decimal	Monto Total de Crédito a Otorgar al Contribuyente
fecha_proceso	Date	Fecha de pago de los montos.
rentas_cos_renta	Integer	Foreing Key de la tabla rentas
ruc_ruc	Decimal	Foreing Key de la tabla ruc

Tabla 7. Ruc

TABLA: Ruc		
Columna	**Tipo de Dato**	**Descripción**
ruc	Decimal	Primary Key del código Ruc de cada Contribuyente
tipo_baja	Integer	Estado que se le da al contribuyente si se encuentra activo o no
desc_dep_mun	Varchar	Descripción del departamento donde reside el contribuyente

Tabla 8. Retenciones

TABLA: retenciones		
Columna	**Tipo de Dato**	**Descripción**
cod_retencion	Integer	Primary Key del código de Retención generado por el Contribuyente
correlativo	Integer	Número secuencial que diferencia de la tabla tenedores a los contribuyentes
periodo	Integer	Indica el valor del impuesto según los tipos de pago.
fecha_emision	Date	Fecha en la que se emitió el pago
concepto	Varchar	Descripción del concepto del pago
monto	Decimal	Monto Inicial del Pago
monto_ir	Decimal	Monto del IR
monto_iva	Decimal	Monto del IVA
crédito_fiscal	Decimal	Monto del Crédito Fiscal otorgado al Contribuyente
fecha_ingreso	Date	Fecha de Inscripción del Contribuyente
retenedores_cod_reporte	Integer	Foreing Key de la tabla retenedores

Tabla 9. Grupos

TABLA: grupos		
Columna	**Tipo de Dato**	**Descripción**
cod_grupo	Integer	Primary Key del código de Grupo del contribuyente
grupo	Varchar	Descripción del grupo que pertenece el Contribuyente: Grande, Pequeño o Cuota Fija
retenciones_cod_retencion	Integer	Foreing Key de la tabla retenciones

Tabla 10. Omisos

TABLA: omisos		
Campo	**Tipo de Dato**	**Descripción**
nombre	Varchar	Nombre del Contribuyente
teléfono	Varchar	Número telefónico del Contribuyente
correo_electronico	Varchar	Correo Electrónico del Contribuyente
id_omisos	Integer	Primary Key del Id de Omiso del contribuyente
retenciones_cod_retencion	Integer	Foreing Key de la tabla retenciones

Tabla 11. Tipos_Documentos

TABLA: tipos_documentos		
Campo	**Tipo de Dato**	**Descripción**
cod_documento	Integer	Primary Key del Codigo de Documento del contribuyente
tipo_documento	Varchar	Código del periodo más un código secuencial por cada contribuyente
retenciones_cod_retencion	Integer	Foreing Key de la tabla retenciones

Tabla 12. Proveedores

TABLA: proveedores		
Campo	**Tipo de Dato**	**Descripción**
cod_proveedor	Integer	Primary Key del Codigo de Proveedor del contribuyente
ruc_proveedor	Decimal	Ruc del Proveedor del Contribuyente
cedula	Varchar	Cédula del Proveedor del Contribuyente
nombre	Varchar	Nombre del Proveedor del Contribuyente
primer_apellido	Varchar	Apellido del Proveedor del Contribuyente
segundo_apellido	Varchar	Segundo Apellido del Proveedor del Contribuyente
nombre_comercial	Varchar	Nombre Comercial del Proveedor del Contribuyente
retenedores_cod_reporte	Integer	Foreing Key de la tabla retenedores

Tabla 13. Detalle_Proveedores

TABLA: detalle_proveedores		
Campo	**Tipo de Dato**	**Descripción**
teléfono	Varchar	Teléfono del Proveedor del Contribuyente
periodo	Integer	Periodo de declaración del Proveedor del Contribuyente
declaración	Integer	Tipo de declaración del Proveedor del Contribuyente
nit_proveedor	Varchar	NIT adscrito del Proveedor del Contribuyente
nombrerep	Varchar	Nombre del Responsable del Proveedor del Contribuyente
concepto	Varchar	Concepto del monto de pago del Proveedor del Contribuyente
monto	Decimal	Monto Total del Proveedor del Contribuyente
monto_ir	Decimal	Monto IR del Proveedor del Contribuyente
monto_iva	Decimal	Monto IVA del Proveedor del Contribuyente
proveedores_cod_proveedor	Integer	Foreing Key de la tabla proveedores

5.3.4. Cuarta Fase: Modelado

La gestión y procesamiento de Big Data es un problema abierto y vigente que puede ser manejado con el diseño de una arquitectura de cuatro niveles, la cual está basada en el análisis de la información y un proceso de modelado para la implementación de la solución del proyecto de Big Data. A continuación en la Ilustración 14, se pueden observar los niveles que contienen un ambiente Big Data y la forma en que se relacionan e interactúan entre ellos:

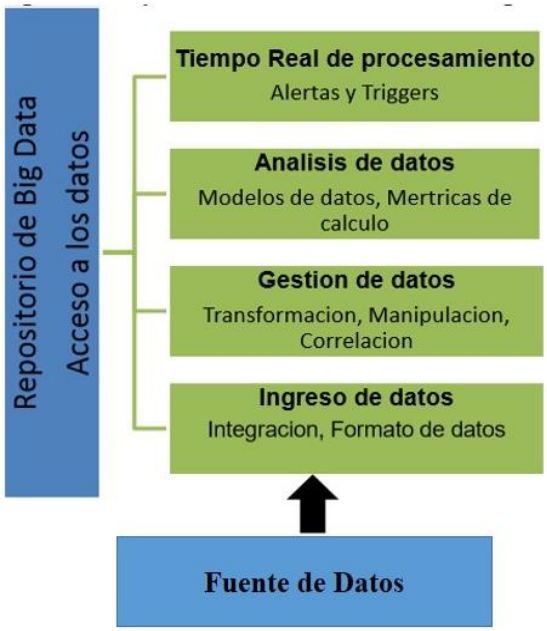

Ilustración 13. Arquitectura de Big Data

A continuación, se detalla en la Ilustración 15, el modelo lógico de la solución:

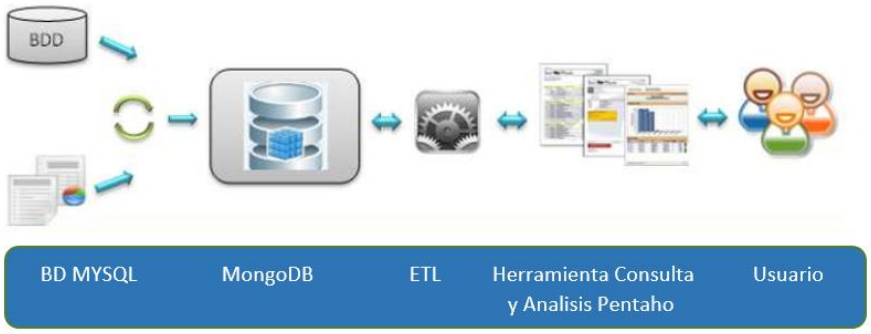

Ilustración 14. Modelo Lógico de la Solución

Se utilizó la herramienta MongoDB, la cual es un complemento de Apache Hadoop para el análisis del modelado de las tablas principales a incluir en la solución del proyecto de Big Data. Se realizó un modelo NoSQL para la base de datos Sairi, con el fin de unificar y lograr un solo planteamiento de almacén de datos para la solución del proyecto del Big Data.

5.3.4.1. MongoDB

MongoDB es un sistema de base de datos NoSQL orientado a documentos. En vez de guardar los datos en tablas, guarda estructuras de datos en documentos tipo JSON con un esquema dinámico, haciendo que la integración de los datos en ciertas aplicaciones sea más fácil y rápida. A continuación en la Tabla 15, se detalla algunas de las características de dicho motor de base de datos NoSQL.

Tabla 14. Especificaciones Técnicas de MongoDB

Desarrollador	10 gen
Última Versión	2.6.0
Estado del desarrollador	Activo
Lenguaje de programación	C++
Sistema operativo	Multiplataforma
Tipo	Base de datos, NoSQL
Licencia	GNU AGPL v3.0
Página web	www.mongodb.com

Esta herramienta ha sido creada para brindar escalabilidad, rendimiento y gran disponibilidad. Aporta un elevado rendimiento, tanto para lectura como escritura, potenciando la computación en memoria. La replicación nativa de MongoDB y la tolerancia a fallos automática ofrece fiabilidad a nivel empresarial y flexibilidad operativa.

Las principales características de esta herramienta son las siguientes:

- Consultas Ad hoc: Soporta la búsqueda por campos, consulta de rangos y expresiones regulares.
- Indexación: Cualquier campo puede ser indexado.
- Replicación: Soporta el modelo maestro esclavo, el maestro puede ejecutar comandos de lectura y escritura, y el esclavo copiar los datos del maestro y utilizarlos para lectura o copia de seguridad, nunca escritura.
- Almacenamiento de archivos: MongoDB puede ser utilizado con un sistema de archivos, gracias a su capacidad de balanceo de carga y replicación de datos utilizando varios servidores.

- Agregación: Se puede utilizar la función MapReduce para procesos batch y operaciones de agregación. Esto permite que los usuarios obtengan un resultado agrupado como en SQL.
- Ejecución de JavaScript del lado del servidor: puede hacer consultas utilizando JavaScript.

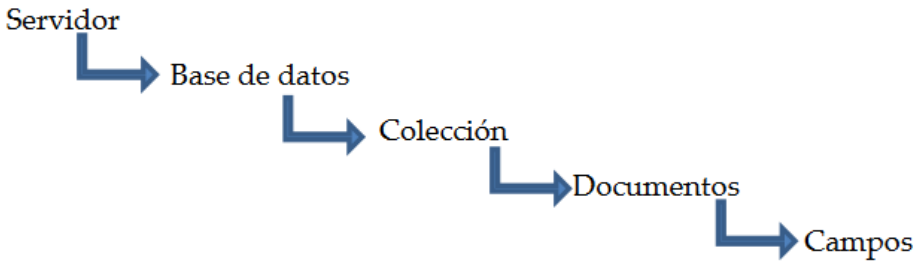

Ilustración 15. Jerarquía de Bases de Datos No SQL orientadas a Documentos

5.3.4.1.1. Modelado en MongoDB

Tipos de datos

A continuación, se detalla algunos de los tipos de datos que soporta MongoDB:

Tabla 15. Tipos de Datos en MongoDB

Tipo de Datos	Descripción
Integer	Números enteros.
Double	Números con decimales.
Boolean	Booleanos verdaderos o falsos.
Date	Fechas.

Timestamp	Estampillas de tiempo.
Null	Valor nulo.
Array	Arreglos de otros tipos de datos.
Object	Otros documentos embebidos.
ObjectID	Identificadores únicos creados por MongoDB al crear documentos sin especificar valores para el campo _id
Data Binaria	Punteros a archivos binarios.
Javascript	Código y funciones Javascript.
String	Cadenas de caracteres.

Patrones de Modelado

Las decisiones para el modelado de los datos implican determinar cómo se deben estructurar los documentos para ganar eficiencia en las consultas y reportes a generarse. Existen dos patrones principales; a continuación se especificara cada uno de ellos.

Embeber

Este patrón se enfoca en incrustar documentos uno dentro de otro con la finalidad de hacerlo parte del mismo registro y que la relación sea directa.

Esta decisión implica la desnormalización de los datos, almacenando dos documentos relacionados en un único documento. Las operaciones a este documento son mucho menos costosas para el servidor que las operaciones que involucran múltiples documentos. En general, se debe de emplear este modelo cuando se tienen relaciones del tipo “contiene” entre entidades.

Una vez explicada una de las formas más común de modelar con la base de datos MongoDB para la solución de nuestra implementación se utilizó dicho patrón de modelado, debido a las relaciones dadas en la base de datos utilizada para el nuevo planteamiento de datos.

Se obtiene finalmente el nuevo modelo de MongoDB deseado, en el cual se debe destacar que fueron utilizadas las tablas y atributos que únicamente son de interés al área de Fiscalización.

Dichas tablas en este nuevo modelo son:

- ✓ Retenedores
- ✓ Retenciones
- ✓ DecSit
- ✓ Departamentos
- ✓ Rentas
- ✓ Ruc
- ✓ Omisos
- ✓ Tipo_Documento

En la Ilustración 17, podemos observar el modelo completo de MongoDB para la implementación de la solución brindada al área de Fiscalización de la Dirección General de Ingresos.

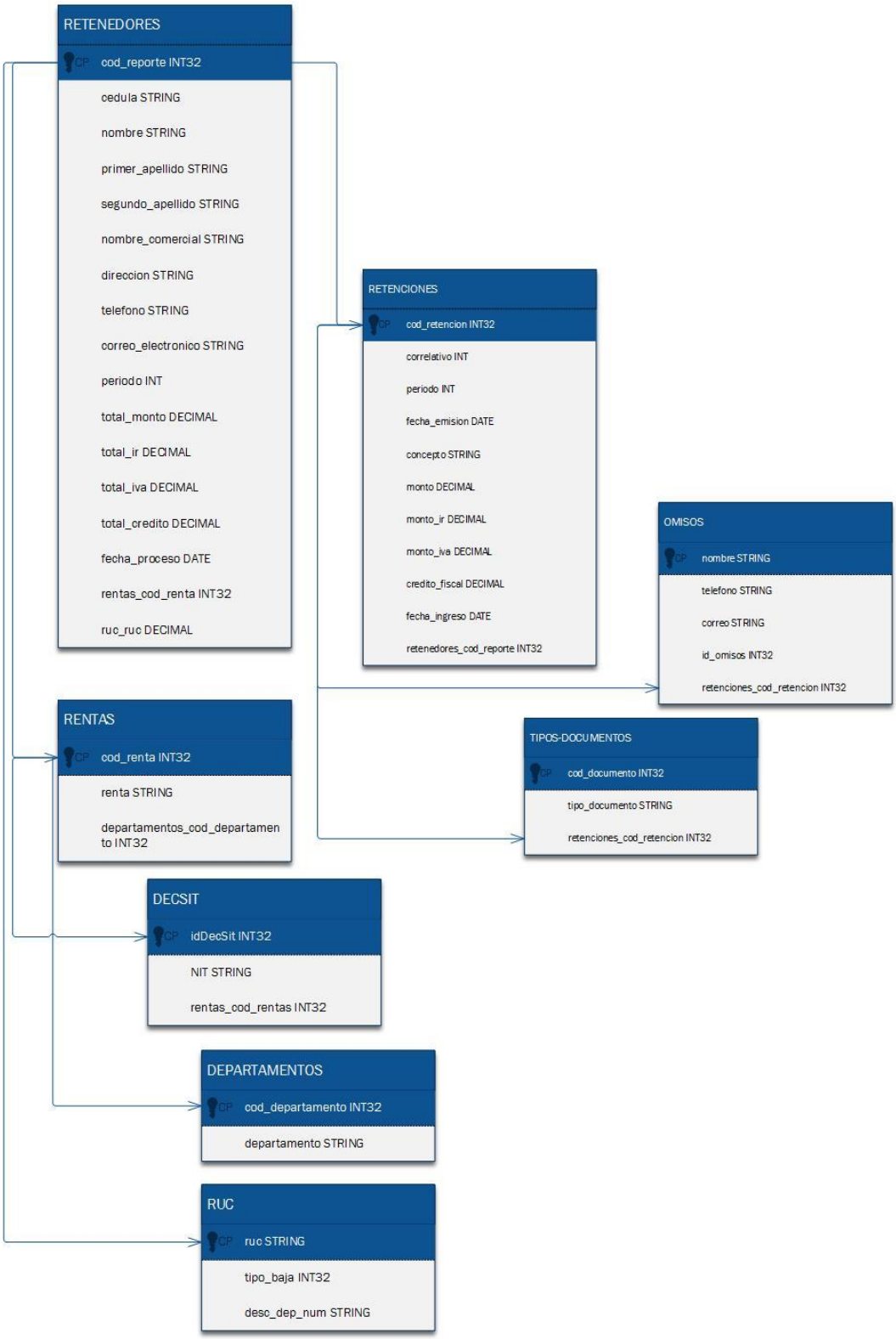

Ilustración 16. Modelo Final MongoDB

5.3.5. Quinta Fase: Evaluación

Como resultado de revisar el proceso del proyecto de la solución de Big Data, se puede fundamentar que la naturaleza cíclica del proceso en la construcción del almacén de datos aumenta su potencialidad. La revisión del proceso también lleva a la institución a comprender los siguientes conceptos:

- Es esencial mantenerse siempre claro en la necesidad de la institución, porque una vez que los datos están preparados para el análisis, es muy fácil iniciar la construcción de modelos.

La institución confía en la precisión y relevancia de los resultados del proyecto y continúa con la fase de desarrollo. Al mismo tiempo, el equipo de proyecto también está listo para la continuación del mismo, lo cual nos llevara a la fase de implantación de la solución de Big Data.

5.3.5.1. Pruebas del Sistema.

En esta etapa se valoraron los logros en cuanto a la calidad de la aplicación realizada y la satisfacción que esta ofrece a los clientes, para ello se analizaron sus factores de éxito a partir de los objetivos y la justificación, encontrándose que para ser exitosa deben realizarse pruebas funcionales, pruebas de comportamiento, pruebas de aceptación.

A continuación, se muestra a detalle cada una de las pruebas antes de la implementación de la solución.

Tabla 16. Detalle de Pruebas de la Solución

Pruebas Funcionales				
Tipo de Pruebas	**Actividades**	**Autores**	**Técnica**	**Resultados**
Prueba de desempeño	Comprobar el tiempo de respuesta para la ejecución de consultas en la carga del ETL	• Arquitecto de Información	Comparar el desempeño del sistema actual con los procedimientos anteriores.	A través de ejecución de transformaciones en el ETL realizo en Pentaho Data Integration se pudo comprobar el tiempo (medido en segundos) de respuesta de la carga de los datos. En la parte de Anexos se verifican los tiempos estimados.
Pruebas de Integridad de Datos y Base de Datos	Asegurar que los métodos de acceso y procesos funcionan adecuadamente y sin ocasionar	• Analista de Información • Arquitecto de Información. • Cliente	• Invocar cada método de acceso y proceso de la Base de datos, utilizando en cada uno datos válidos e inválidos.	Se comprobó al realizar consultas entre MySQL y MongoDB, verificando así que la integridad de los datos es la misma.

	corrupción de datos.		• Analizar la Base de datos, para asegurar que los datos han sido grabados apropiadamente, que todos los eventos de Base de datos se ejecutaron en forma correcta y revisar los datos retornados en diferentes consultas.	
Pruebas de Seguridad y Control de Acceso	Verificar que un actor solo pueda acceder a las funciones y datos que su usuario tiene permitido.	• Arquitecto de Información	• Comprobar la seguridad del sistema, incluyendo acceso a datos o Funciones de negocios • Comprobar la seguridad de la Base de Datos, incluyendo	Una única conexión al Servidor realizada por el Arquitecto de Información.

			ingresos y accesos remotos al sistema.	
Pruebas de Comportamiento				
Pruebas de Stress	Verificar que el sistema funciona apropiadamente y sin errores, bajo condiciones de stress	• Arquitecto de Información • Cliente • Usuario	• ¿Memoria baja o no disponible en el servidor? • ¿Múltiples usuarios desempeñando la misma transacción con los mismos datos?	La implementación del aplicativo se encuentra instalado en un servidor de Pruebas con las mismas características del servidor de producción, además se utilizó la herramienta: Nagios para el monitoreo continuo de cada uno de los servicios instalados en el servidor. En los anexos se adjuntan las imágenes correspondientes de las pruebas.

Pruebas de Volumen	Las pruebas de volumen hacen referencia a grandes cantidades de datos para determinar los límites en los cuales el Sistema falle.	• Arquitecto de Información	• Determinar el Máximo tamaño de la base de datos (actual o escalado) y múltiples consultas ejecutadas simultáneamente	El motor de Base de Datos NoSQL de MongoDB utiliza una serialización binaria de JSON, llamada BSON; el cual posee un tamaño especifico de 4MB por documento. En la parte de anexos detallaremos como están compuestos cada uno de los documentos de la Base de Datos. Ver, ANEXO C: Pruebas de Volumen.
Pruebas de Aceptación				
Pruebas del Ciclo del Negocio	Asegurar que la solución funciona de acuerdo con el modelo de negocios emulando	• Cliente	• Todas las funciones ocurren en un periodo de tiempo serán ejecutadas en el tiempo apropiado.	Análisis de Reportería cumpliendo con los requerimientos establecidos.

	todos los eventos en el tiempo y en función del tiempo.		• Cada regla de negocios es aplicada adecuadamente.	Ver, ANEXO B: Pruebas del Ciclo del Negocio.
Pruebas de Configuración	Validar y verificar que la solución funciona apropiadamente en las estaciones de trabajo recomendadas.	• Analista de Información • Arquitecto de Información	• Incluir la apertura o cierre de varias aplicaciones diferentes a las utilizadas para la solución (o algún tipo de software similar a la que se está probando) como una parte de la prueba, ya sea al comienzo o en algún momento intermedio.	Las configuraciones de los servicios utilizados para la solución se detallan en la parte de anexos y las debidas pruebas de los mismos.
Prueba de Aceptación	Determinar por parte del cliente la aceptación o rechazo de la	• Cliente • Usuario	• La prueba de aceptación es ejecutada antes de que la aplicación sea	La solución planteada está siendo utilizada en un Ambiente de

	solución implementada.		instalada dentro de un ambiente de producción.	Desarrollo para las debidas pruebas.
Prueba de Instalación	Verificar y validar que la solución se instala apropiadamente en cada cliente.	• Arquitecto de Información • Cliente • Usuario	• Instalaciones nuevas, nuevas máquinas a las que nunca se les ha instalado la solución.	Se instaló y configuro de forma correcta y rápida el sistema en los nuevos servidores. Se instaló y configuro de forma correcta y rápida el sistema en servidores ya en uso.

Además, se detalla los requerimientos funcionales (Reportes) solicitados por los usuarios del área de fiscalización.

Tabla 17. Descripción de Requerimientos Funcionales (Reportes)

Nombre Reporte	Título Reporte	Descripción	Filtro
Dirección General de Ingresos – Área de Fiscalización	Contribuyentes Omisos por Renta	• Permite visualizar los contribuyentes de una zona determinada (por Renta), en el cual se podrá observar los siguientes datos de los contribuyentes: Nombre, Primer_Apellido, Nombre_Comercial, Ruc, Total_IR, Renta. • Se detalla un dashboards en cual nos indicara de manera gráfica que Renta genera más declaraciones en un periodo determinado.	• Renta
Dirección General de	Contribuyentes Omisos por	• Permite visualizar los contribuyentes omisos categorizados por un	• Tipo Documento

Ingresos – Documentos	Tipo de Documento	tipo de Documento (por Renta), en el cual se podrá observar los siguientes datos de los contribuyentes: Nombre, Ruc, Tipo_Documento, Fecha_Ingreso	
Dirección General de Ingresos - Omisos	Contribuyentes Omisos Totales	• Permite visualizar todos los contribuyentes a nivel nacional, en el cual se podrá observar los siguientes datos de los contribuyentes: Nombre_Comercial NIT, Direccion, Cedula, Monto_IR, Total_IVA, Total_IR, Fecha_Emision	• Año

Para dar respuesta a los requerimientos funcionales solicitados por los usuarios se realizó una encuesta a los funcionarios del área de la División de Informática y Sistemas de la DGI y verificar que la nueva solución de Big

Data implementada cumple con los requerimientos funcionales que se desean obtener.

5.3.6. Sexta Fase: Implementación

La implantación está contemplada en cumplir con cada una de las tareas en tiempo y forma, la comunicación constante entre todos los implicados del proyecto construyo un pilar fundamental para la conclusión de cada fase en todo el ciclo del proyecto.

En la entrega final del proyecto, se realizó un informe final en el cual se contempla a detalle cada fase del proyecto.

5.3.6.1. Estudio De Costos

Partiendo de la premisa que el proyecto se implementó dentro de una institución del estado ya consolidada y con la División de Informática en perfecto funcionamiento, los costos generados adicionales por este proyecto se detallan en: Mano de Obra(Costos Directos) y Materia Prima(Costos Directos).

Tabla 14. Mano de Obra - Costos Directos – Elaboración Propia

Mano de Obra Costos					
Cargo	**Tiempo**	**Salario Mensual**	**Otros / mensual**		**Total**
Analista de Información (Trabajador de la Empresa)	3 meses	$ 0.000 (La institución asigno el apoyo de este recurso)	Alimentación y Viáticos	$100	$ 300.00
Usuarios Finales (Personal del Área de Fiscalización de la DGI)	3 meses	$ 0.000 (La institución asigno el apoyo de este recurso)	Alimentación y Viáticos	$ 30	$ 90.00
Auxiliar de Pruebas	1 meses	$ 0.000 (La institución asigno el apoyo de este recurso)	Alimentación y Viáticos	$ 50	$ 50.00
Costo Total Inversión en Recurso					**$440.00**

Tabla 15. Materia Prima - Costos Directos – Elaboración Propia

Materia Prima Costos				
Materia Prima	**Cantidad**	**Valor**	**Otros**	**Total**
Pc's ambiente de desarrollo	2	$0.00	-	$0.00
Servidor de Producción.	1	$0.00	-	$0.00
Licencia MongoDB (Para un servidor)	1	$ 500	-	$ 500.00
Licencia Pentaho (Para un servidor)	2	$950	-	$950.00
Total Costos Materia Prima				**$1450.00**

6. CONCLUSIONES Y RECOMENDACIONES

- ✓ Se logró implementar una solución basada en Big Data para practicar las técnicas de análisis, que faciliten el seguimiento y control en la información tributaria del área de Fiscalización en la Dirección General de Ingresos.

- ✓ La utilización de herramientas como Pentaho y MongoDB permitió la construcción de un ETL para enriquecer un sistema de reportería logrando resultados positivos en la presentación de los datos a través de reportes dinámicos y tableros.

- ✓ La evaluación realizada de la solución de Big Data permitió verificar que dicha solución es adecuada y pertinente para los objetivos propuestos, permitiendo importantes ahorros de tiempo en los procesos de obtención y análisis de información, además de ser más fácil de usar que el método anterior.

- ✓ La Metodología de CRISP resultó ser muy completa y apropiada para guiar el proceso de implementación de la solución de Big Data.

Algunas de las recomendaciones brindadas sobre la solucion de Big Data al Centro Nacional de Datos Fiscales son:

- ✓ Capacitar al personal de la Unidad de Bases de Batos y Sistemas Operativos sobre el correcto funcionamiento de las soluciones de BIG DATA, para la migración de las Bases de Datos SQL a NoSQL.

- ✓ Elaborar documentación técnica sobre los futuros proyectos de BIG DATA, tales como diseño de la estructura, justificación de su planteamiento y descripción de las funcionalidades.

7. BIBLIOGRAFIA

Acens. (s.f.). *Acens company.* Obtenido de Acens company: https://www.acens.com/wp-content/images/2014/02/bbdd-nosql-wp-acens.pdf

Dev, G. (s.f.). *NoSQL: clasificación de las bases de datos según el teorema CAP*. Obtenido de https://www.genbetadev.com/bases-de-datos/nosql-clasificacion-de-las-bases-de-datos-segun-el-teorema-cap

Foundation, T. A. (21 de Mayo de 2017). *Welcome to Apache Avro!* Obtenido de Welcome to Apache Avro!: http://avro.apache.org/

Fragoso, R. B. (18 de 06 de 2012). *¿Qué es Big Data? Todos formamos parte de ese gran crecimiento de datos.* Obtenido de ¿Qué es Big Data? Todos formamos parte de ese gran crecimiento de datos: https://www.ibm.com/developerworks/ssa/local/im/que-es-big-data/

Garcete, A. (s.f.). *Base de Datos Orientado a Columnas.* Asunción - Paraguay: Universidad Catolica "Nuestra Señora de la Asunción".

García, D. L. (2012). *Análisis de las posibilidades de uso de Big Data en las organizaciones.* -: -.

GARTNER. (22 de Septiembre de 2013). *Big Data. Connecticut: Gartner*. Obtenido de www.gartner.com/it-glossary/big-data

Gravitar, I. S. (s.f.). *Pentaho* . Obtenido de Pentaho : http://gravitar.biz/

IBM. (2012). *Manual CRISP-DM de IBM SPSS Modeler.* Obtenido de ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/15.0/es/CRISP-DM.pdf

LópMor, E. (29 de Abril de 2011). *Manual Pentaho Report Designer*. Obtenido de https://es.scribd.com/document/54202254/Manual-Pentaho-Report-Designer

Services, I. (2012). *Analytics: el uso de big data en el mundo real: Cómo las empresas más innovadoras extraen valor de datos inciertos.* Madrid, España: Santa Hortensia, 26-28 28002.

Silberschatz, A. (2002). FUNDAMENTOS DE BASES DE DATOS. En H. F. Abraham Silberschatz, *FUNDAMENTOS DE BASES DE DATOS* (pág. 1). España: Concepción Fernández Madrid.

Soubra, D. (05 de Julio de 2012). *Data Sciencie Central*. Obtenido de Data Sciencie Central: http://www.datasciencecentral.com/forum/topics/the-3vs-that-define-big-data

8. GLOSARIO DE TÉRMINOS

Ad Hoc: análisis de datos en donde hay una flexibilidad lo más amplia posible en cuanto a los formatos, consultas predefinidas, valores preseleccionados, etc.

CNDF: Centro Nacional de Datos Fiscales.

CRISP: Cross Industry Standard Process for Data Mining.

ETL: Extract, Transform and Load.

Exabytes: Un exabyte (abreviado "EB") es una unidad de medida de almacenamiento de datos que equivale a 10^{18} bytes.

GPS: Sistema de Posicionamiento Global.

JSON: JavaScript Object Notation, es un formato ligero de intercambio de datos.

MapReduce: framework que proporciona un sistema de procesamiento de datos paralelo y distribuido.

NoSQL: es un término que describe las bases de datos no relacionales de alto desempeño. Las bases de datos NoSQL utilizan varios modelos de datos, incluidos los de documentos, gráficos, claves-valores, columnas, etc.

Omisos: Referencia al estado de pagos de los contribuyentes.

PDI: Pentaho Data Integration.

Petabytes: Se trata de una unidad más grande que el gigabyte o el terabyte, pero más pequeña que unidades como el exabyte, el zettabyte o el yottabyte.

PRD: Pentaho Report Designer.

Rentas: Referencia a los distintos lugares localizados en Managua para el pago de impuestos a la DGI.

Reportería: es un informe o una noticia. Este tipo de documento (que puede ser impreso, digital, audiovisual, etc.) pretende transmitir una información, aunque puede tener diversos objetivos.

Retenciones: Referencia a los pagos realizados a los contribuyentes.

Ruc: ID único para identificar a contribuyentes.

Tributos: son los aportes que todos los contribuyentes tienen que transferir al Estado.

XML: es un subconjunto de SGML (Estándar Generalised Mark-up Language), simplificado y adaptado a Internet.

9. ANEXOS

9.1. ANEXO A: Encuesta Funcionarios al Área de División de Informática y Sistemas de la DGI

Para identificar fácilmente los resultados obtenidos al área de Fiscalización se realizó una encuesta, a continuación se describe la pregunta y se analizaron las respuestas como se muestran a continuación:

1. ¿Qué bases de datos ha adquirido/ usado la organización para el almacenamiento de información en los últimos 5 años?
   R= De acá podemos comprobar que nunca se han utilizado bases de datos NoSqL en la institución.

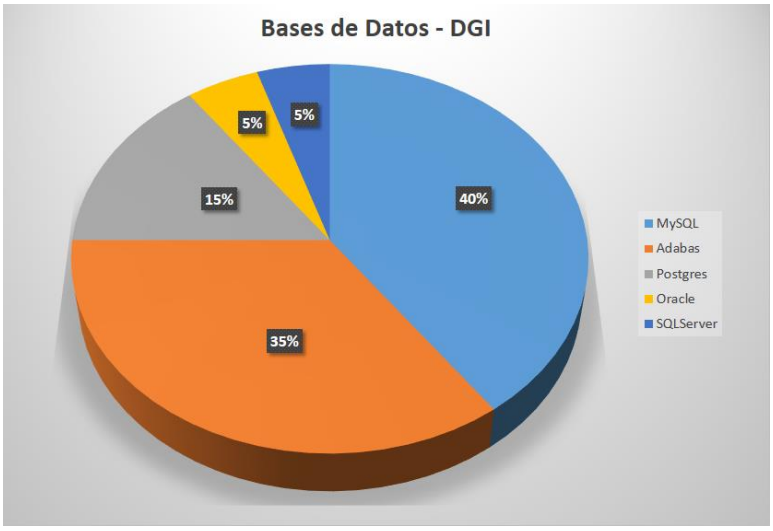

Ilustración 17. Respuesta #1 de la Encuesta

2. ¿Qué factores se tomaron en cuenta para la elección de las Bases de Datos que se utilizan en la DGI?

Acá se definen las ventajas que tenemos con mongodb en comparación a las bases de datos que estamos utilizando, ya que la base de datos NoSQL implementada nos da más tiempo de respuesta a la solicitud de los requerimientos que fueron detallados.

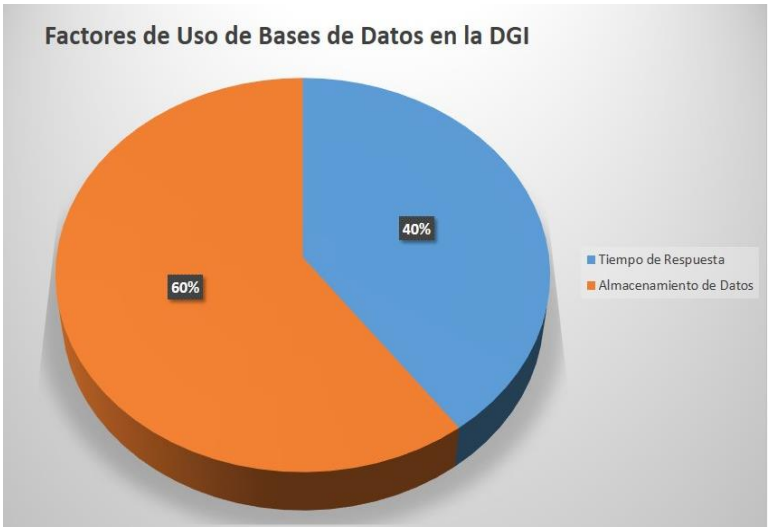

Ilustración 18. Respuesta #2 de la Encuesta

3. ¿Sobre qué sistema operativo se ejecutan los servidores de bases de datos?

   Se aseguró que la solución implementada se ejecute bajo un entorno de producción en un sistema operativo GNU/Linux ya que es el más utilizado por la institución.

Ilustración 19. Respuesta #3 de la Encuesta

4. ¿Las aplicaciones actuales son en su mayoría desarrollos internos o realizados por terceros?

   La DGI en su mayoría realiza aplicaciones con desarrollo interno, es por esta restricción que el desarrollo de la solución debe ser únicamente para uso de la institución y destinada al área de fiscalización.

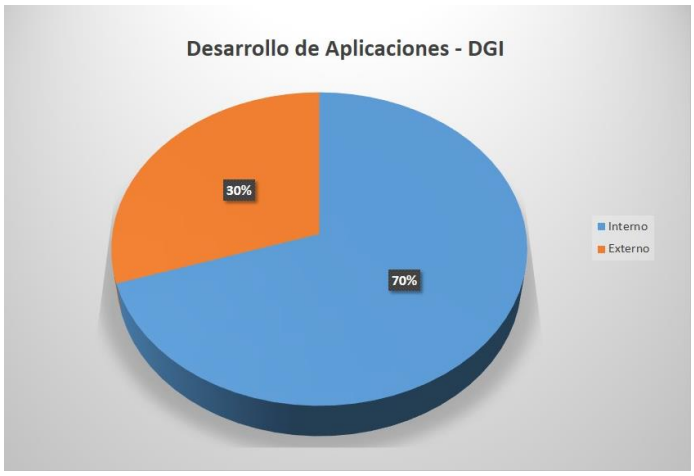

Ilustración 20. Respuesta #4 de la Encuesta

5. ¿Conoce el termino NoSQL?

   Se determina que el término NoSQL no es muy conocido por la División de Informática de la DGI, por lo cual se propuso una solución nueva e integral para dar respuesta a los requerimientos solicitados.

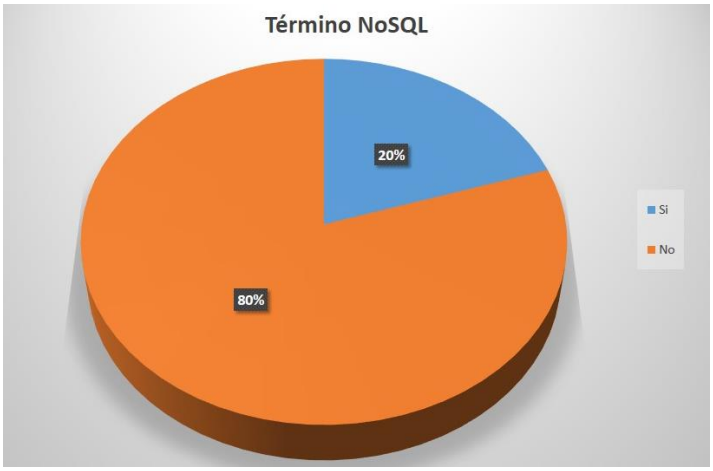

Ilustración 21. Respuesta #5 de la Encuesta

6. De forma general, ¿Qué consideraciones tienen presentes para la migración de los datos a una nueva versión u otro motor de base de datos orientados a NoSQL?

   Se comprobó que la Base de Datos NoSQL orientada a Documentos (MongoDB), cumple con cada una de las consideraciones para la migración y de esta forma desarrollar una solución que cumpla con los requisitos solicitados.

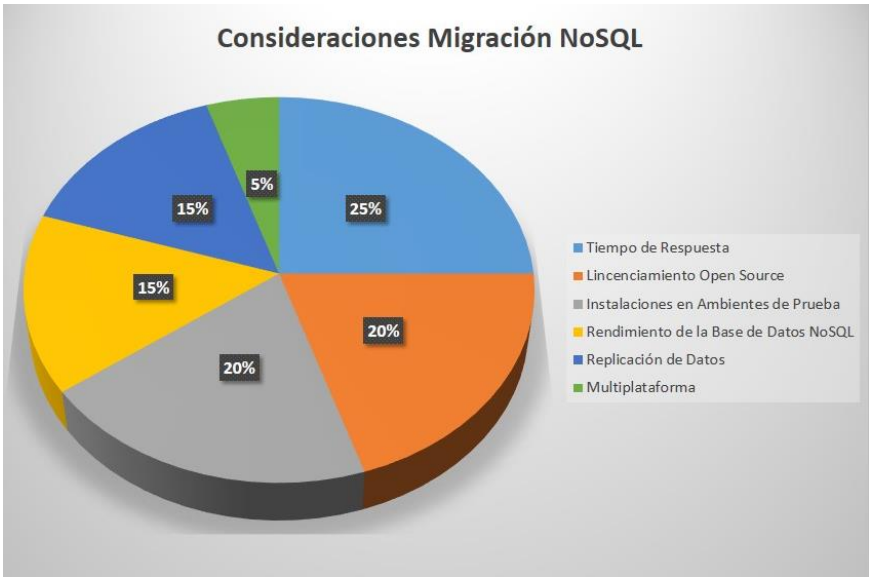

Ilustración 22. Respuesta #6 de la Encuesta

7. Actualmente, ¿existen reportes interactivos para el manejo de la información en tiempo y forma según los requerimientos solicitados?
   En el levantamiento de requerimientos se determina que no existen reportes interactivos para el manejo de la información y que estos reportes se generaban en Excel y los funcionarios encargados de hacer la revisión y consulta de dicha información se tardaban demasiado tiempo, de esta forma podemos comprobar que la nueva solución implementada brinda mayor agilidad en las consultas y revisiones de dicha información.

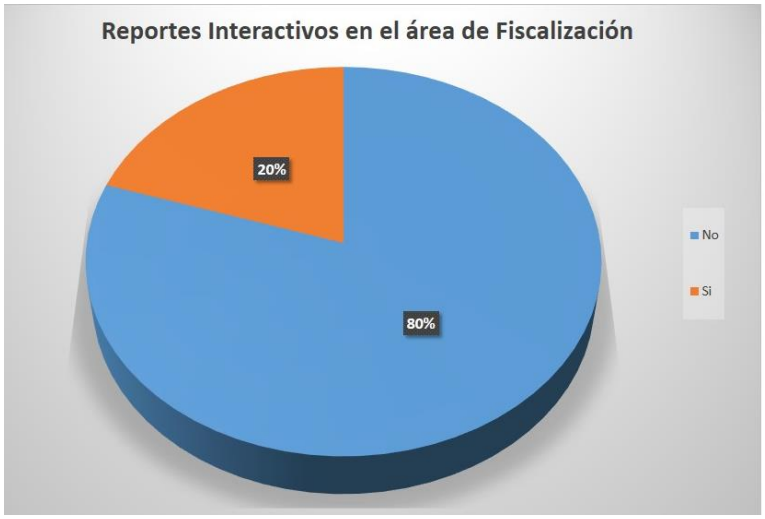

Ilustración 23. Respuesta #7 de la Encuesta

8. ¿Con los reportes presentados al área de fiscalización se tiene mayor visualización de los datos?

   Se determina que la solución brindada al área de Fiscalización consiste en la realización de 4 reportes definimos como: Contribuyentes Omisos por Renta, Contribuyentes Omisos por Tipo de Documento, Contribuyentes por omisos Totales, dan una mejor experiencia al usuario en la visualización y manejo de la información de los contribuyentes.

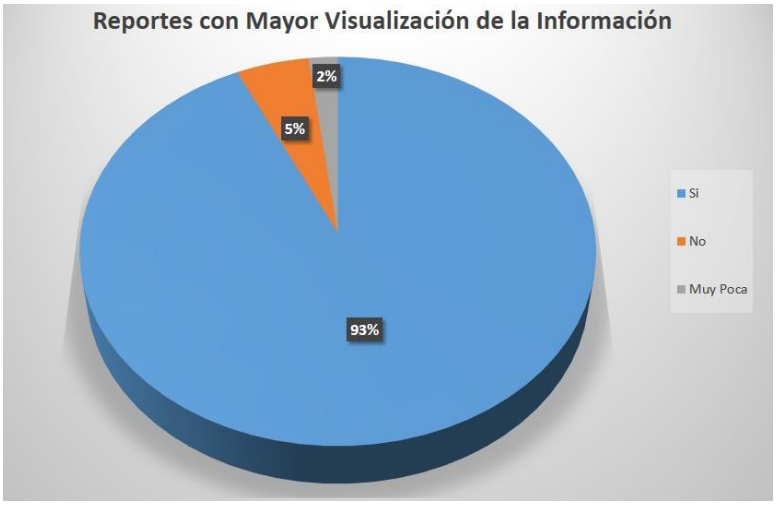

Ilustración 24. Respuesta #8 de la Encuesta

9.2. ANEXO B: Pruebas del Ciclo del Negocio

Para poder evaluar los resultados de la aplicación de la herramienta de Big Data y poder verificar si cumple con los objetivos planteados para la DGI, se analizaron sus factores de éxito a partir de los objetivos y la justificación, encontrándose que para ser exitosa debe:

- Requerir menos reportes.
- Requerir menos tiempo para realizar la consulta.
- Ser más fácil de usar.
- Ser interactiva.

9.2.1. Definición de las métricas

A partir de analizar la naturaleza de los factores de éxito y agruparlos de acuerdo a éstas, se determinaron las métricas y medidas a utilizar:

- ✓ Desempeño:

Tiempo de respuesta

Numero de reportes

- ✓ Facilidad de uso (usabilidad):
  - Interfaz simple e intuitiva
  - Es Interactiva
  - Tiene información visual

Las medidas de desempeño se obtuvieron a través de la aplicación de un test comparativo con las herramientas actuales (reportes tabulares) y la solución de Big Data desarrollada, en el cual se cronometró el tiempo y se contaron los números de reportes (o consultas) diferentes para obtener la información requerida.

Las medidas de usabilidad se obtuvieron mediante instrumento de encuesta a los testers de su percepción respecto al uso de las herramientas.

9.2.2. Definición de instrumentos

La encuesta consultó para cada herramienta una única respuesta para cada medida por testers, de acuerdo a las siguientes escalas de valores:

Tabla 16. Definición de Instrumentos - Elaboración Propia

Medida de evaluación	Escalas de valores
Interfaz simple e intuitiva	Simple, Compleja
Es Interactiva	Sí, No
Tiene información visual	Sí, No

9.2.3. Resultados de la evaluación

A continuación se presentan los resultados de ambas evaluaciones: Test para comparar el desempeño de las herramientas y encuesta para evaluar la percepción sobre la usabilidad de la nueva herramienta en comparación con la anterior.

Se aplicó el test por parte de funcionarios de nivel táctico del órgano rector de la DGI. El equipo de testers consistió en cinco funcionarios (fiscalizadores), realizando los casos de prueba tanto con la herramienta anterior como con la nueva.

9.2.3.1. Resultado del Caso de Prueba 1: Reporte Contribuyentes Omisos por Renta

Verificar información de los Contribuyentes Omisos por Renta.

$$Ahorro = \frac{(Resultado_{metodo_{anterior}} - Resultado_{BI})}{Resultado\_metodo\_anterior} * 100\%$$

Tabla 17. Caso de Prueba 1 - Elaboración Propia

C1	Herramienta Anterior (A)		Solución Big Data (B)		Ahorro (A - B)/A*100%	
Tester	**Tiempo de Respuesta (s)**	**# de Reportes**	**Tiempo de Respuesta (s)**	**# de Reportes**	**Tiempo (%)**	**# de reportes (%)**
1	286	1	187	1	35%	0%
2	301	1	199	1	34%	0%
3	256	1	179	1	30%	0%
4	277	1	187	1	32%	0%
5	294	1	194	1	34%	0%
Media	283	1	189	1	33%	0%

Según se puede observar en la Tabla 17, en este caso se obtuvo un ahorro de tiempo promedio del 33%. No hubo ahorro en cantidad de pasos procedimentales.

9.2.3.2. Resultado del Caso de Prueba 2: Reporte de Contribuyentes Omisos por Tipo de Documento

Verificar información de los Contribuyentes Omisos por Tipo de Documento (097 y 037).

Tabla 18. Caso de Prueba 2 - Elaboración Propia

C2	Herramienta Anterior (A)		Solución Big Data (B)		Ahorro (A - B)/A*100%	
Tester	**Tiempo de Respuesta (s)**	**# de Reportes**	**Tiempo de Respuesta (s)**	**# de Reportes**	**Tiempo (%)**	**# de reportes (%)**
1	564	2	282	1	50%	67%
2	593	2	266	1	55%	67%
3	532	2	245	1	54%	67%
4	597	2	271	1	55%	67%
5	578	2	263	1	54%	67%
Media	573	2	265	1	54%	67%

Según se puede observar en la Tabla 18, en este caso se obtuvo un ahorro de tiempo promedio del 54%. Asimismo se ganó eficiencia en un 67% de ahorro en cantidad de pasos procedimentales.

9.2.3.3. Resultado del Caso de Prueba 3: Reporte de Contribuyentes por Omisos Totales

Verificar información de los Contribuyentes por Omisos Totales

Tabla 19. Caso de Prueba 3 - Elaboración Propia

C3	Herramienta Anterior (A)		Solución Big Data (B)		Ahorro (A - B)/A*100%	
Tester	**Tiempo de Respuesta (s)**	**# de Reportes**	**Tiempo de Respuesta (s)**	**# de Reportes**	**Tiempo (%)**	**# de reportes (%)**
1	387	3	220	1	43%	50%
2	394	3	234	1	41%	50%
3	399	3	245	1	39%	50%
4	385	3	235	1	39%	50%
5	390	3	241	1	38%	50%
Media	391	3	235	1	40%	50%

Según se puede observar en la Tabla 19, en este caso, se obtuvo un ahorro de tiempo promedio del 97%. Se obtuvo un ahorro de 50% en cantidad de pasos procedimentales.

9.2.3.4. Resultados de los Casos 1,2 y 3

Se aplicó la encuesta a los funcionarios testers. Esta consulta consistió en tres preguntas (medidas de usabilidad) que se preguntaron tanto para el método anterior como para la nueva herramienta, obteniéndose consenso que el método anterior es complejo y la solución de Big Data implementada es más simple e intuitiva y por ende, más fácil de usar. Asimismo, mientras el método anterior carece de interactividad para el análisis y ayudas visuales, la nueva solución de Big Data sí lo tiene.

Tabla 20. Resultado de Casos 1,2 y 3 - Elaboración Propia

Encuesta	Interfaz simple e intuitiva		Es Interactiva		Tiene información visual	
Tester	**Método Anterior**	**Solución Big Data**	**Método Anterior**	**Solución Big Data**	**Método Anterior**	**Solución Big Data**
1	Compleja	Simple	NO	SI	NO	SI
2	Compleja	Simple	NO	SI	NO	SI
3	Compleja	Simple	NO	SI	NO	SI
4	Compleja	Simple	NO	SI	NO	SI
5	Compleja	Simple	NO	SI	NO	SI
Moda	Compleja	Simple	NO	SI	NO	SI

En conclusión, los resultados de la evaluación evidencian que la solución de Big Data implementada, va a proveer todos los beneficios esperados a los usuarios de la DGI, lográndose así los objetivos del proyecto exitosamente.

9.3. ANEXO C: Pruebas de Volumen

El motor de Base de Datos NoSQL de MongoDB utiliza una serialización binaria de JSON, llamada BSON; el cual posee un tamaño especifico de 4MB por documento.

A continuación en la Ilustración 26, se muestra la estructura de un documento completo en MongoDB.

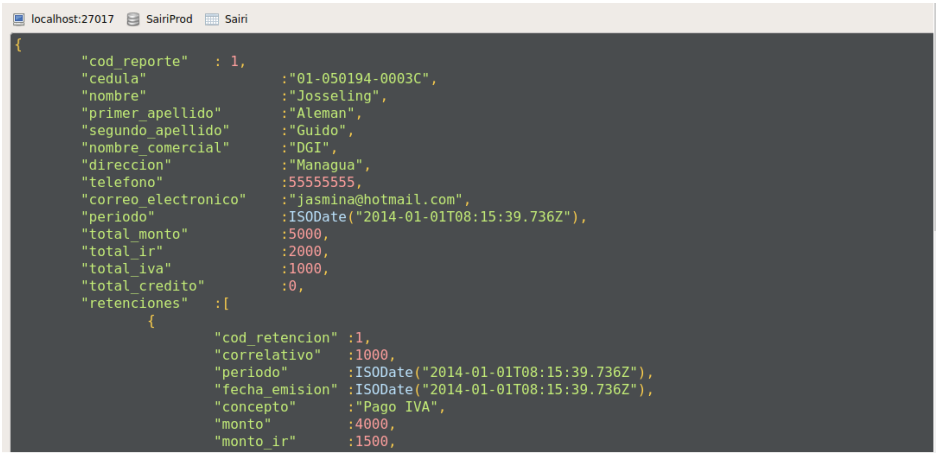

Ilustración 25. Documento Formato JSON

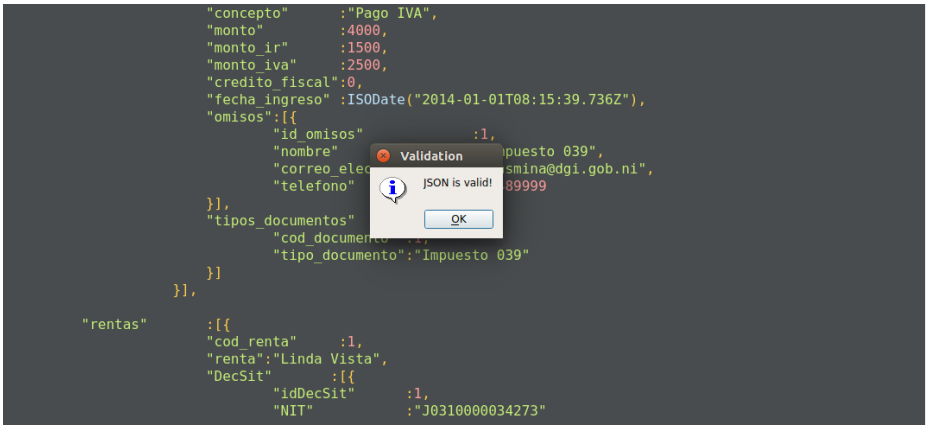

Ilustración 26. Validación de la Inserción del Formato JSON

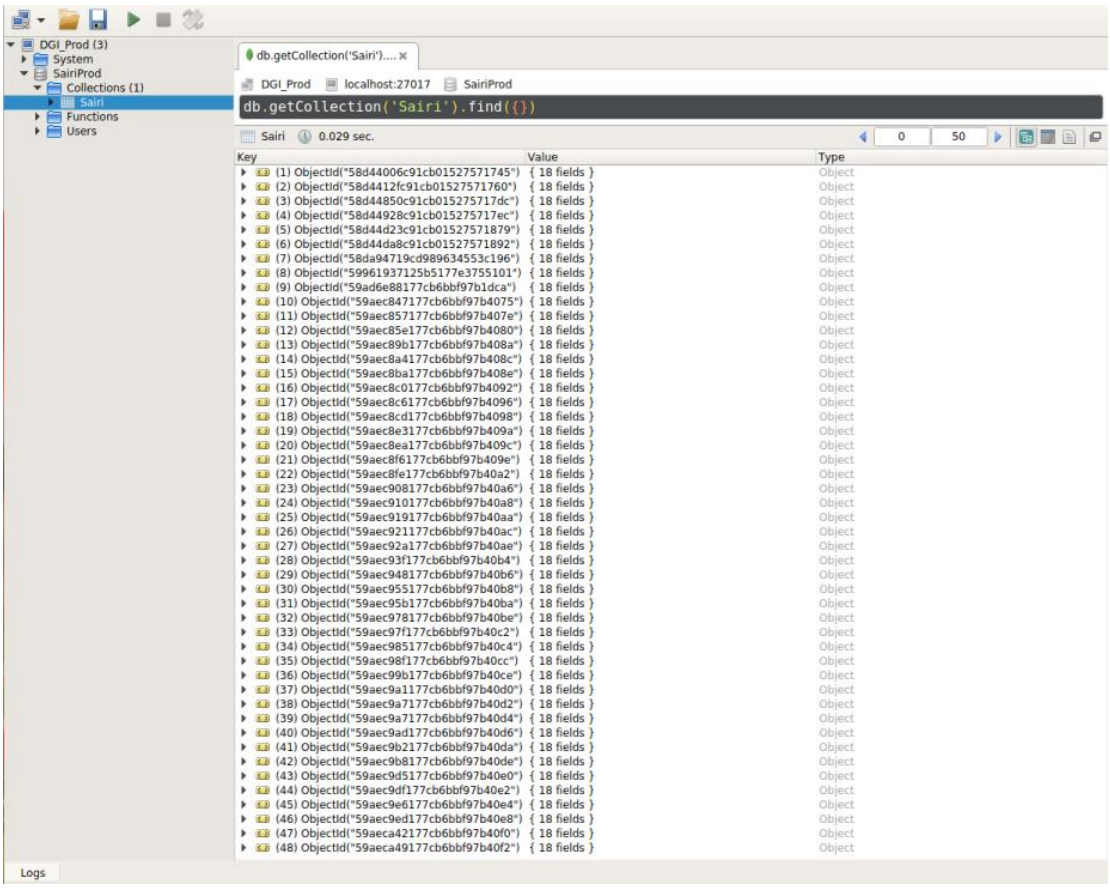

Ilustración 27. Colecciones de MongoDB en Formato Object

9.4. ANEXO D: Implementación de la Solución de Big Data

Para dicha implementación se utilizaron dos servidores (Prueba y Producción), los cuales fueron proporcionados por la División de Informática y Sistemas de la Dirección General de Ingresos.

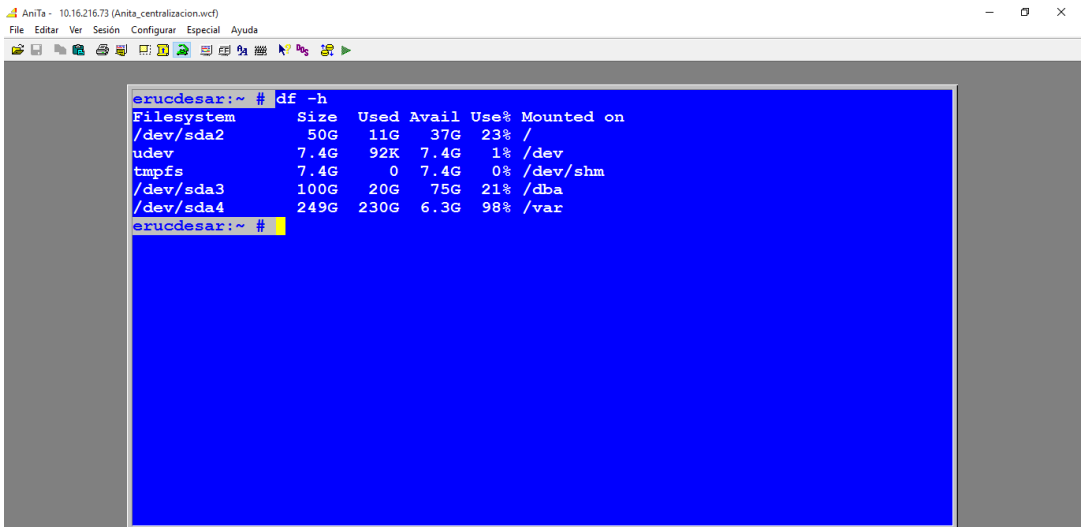

Ilustración 28. Servidor para Pruebas - Desarrollo

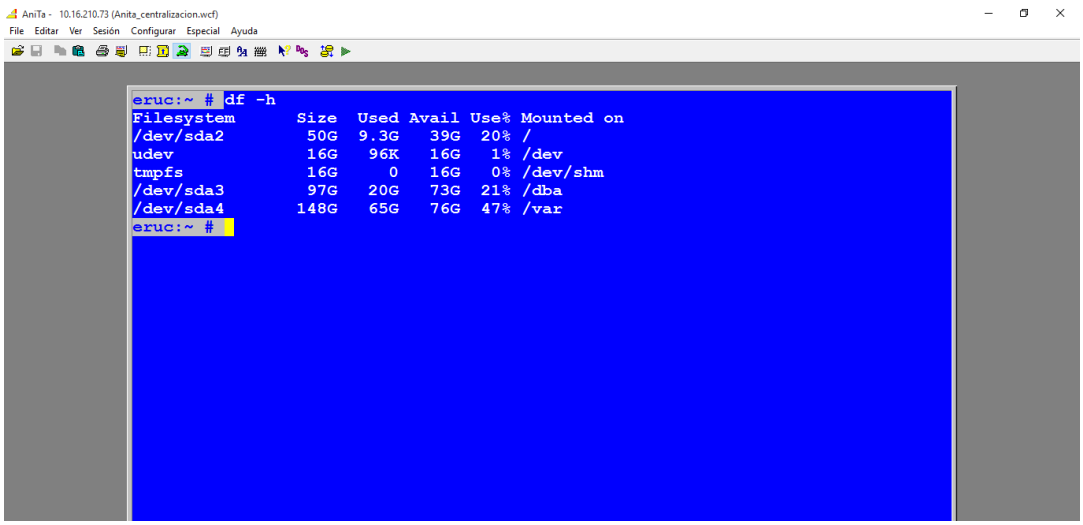

Ilustración 29. Servidor de Producción

Se realizó la configuración de MongoDB en ambos servidores para las debidas pruebas. Para una mejor comprensión del sistema de MongoDB se utilizó un gestor para dicha Base de Datos; llamado RoboMongo

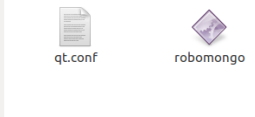

Ilustración 30. Ícono RoboMongo

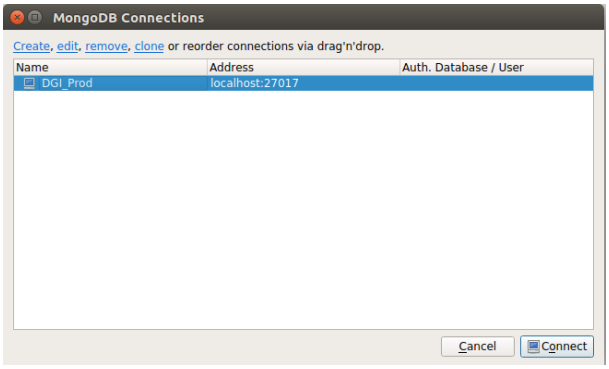

Ilustración 31. Conexión al Servidor RoboMongo

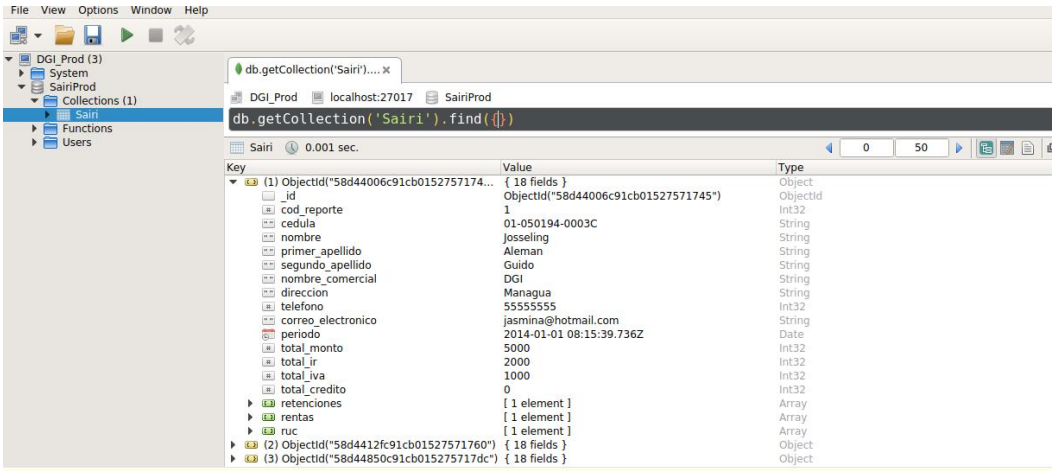

Ilustración 32. Pantalla Inicial RoboMongo

Ahora se detalla el proceso de Integración de los Datos, utilizando la herramienta Pentaho Data Integration.

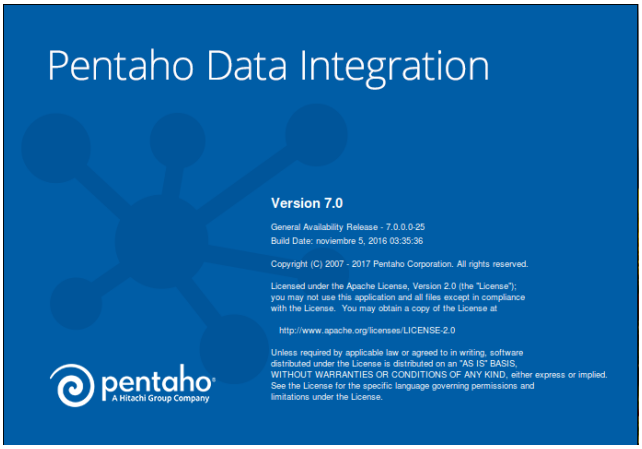

Ilustración 33. Iniciando Pentaho Data Integration

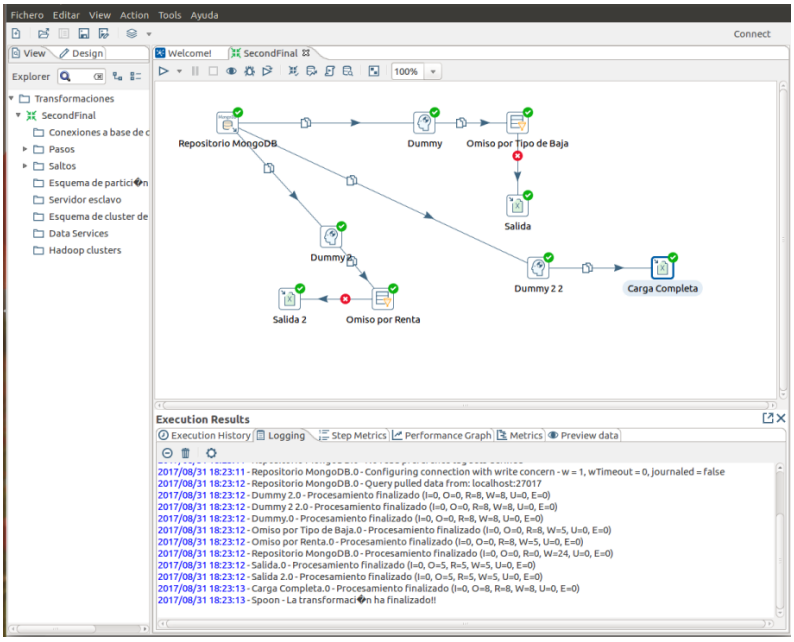

Ilustración 34. Modelo Final - Pentaho Data Integration

Por último se detalla la solución final, utilizando la herramienta Pentaho Data Report.

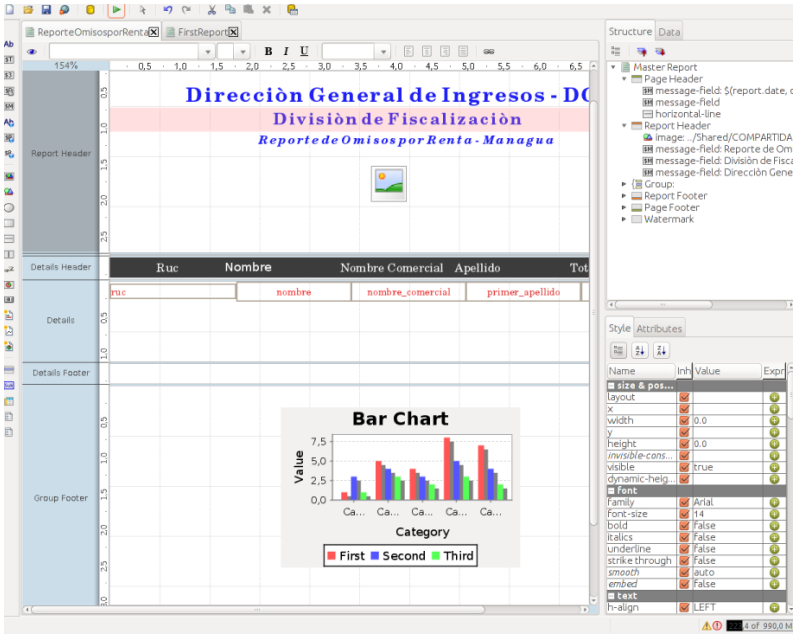

Ilustración 35. Reporte en Pentaho Data Integration

Monitoreo de Nagios para Servidor de Pruebas y Producción

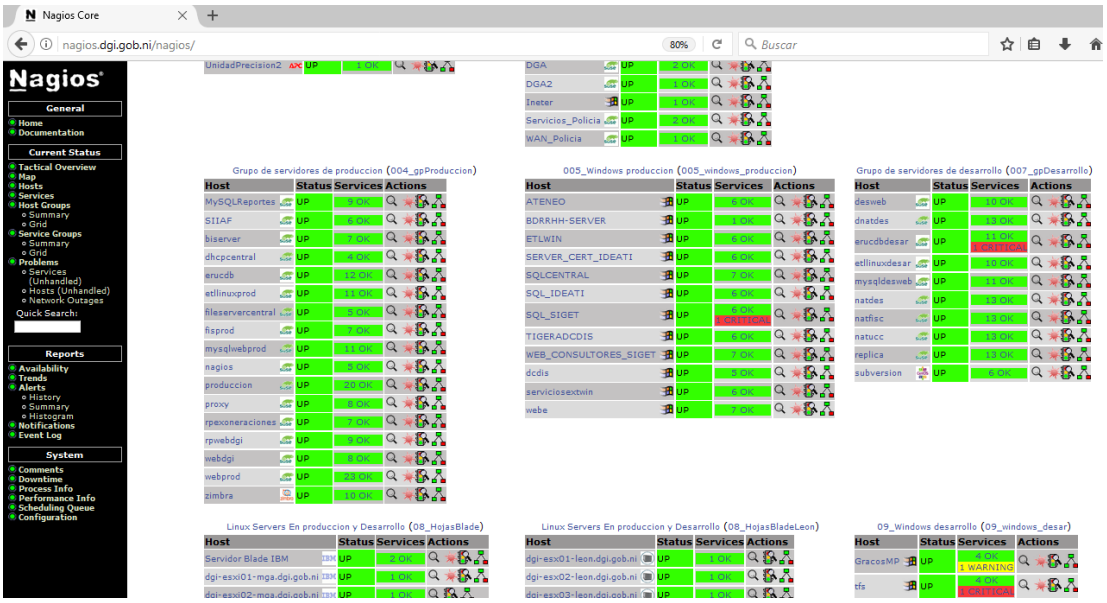

Ilustración 36. Monitoreo Servidores Nagios

9.5. ANEXO E: Nota Aclaratoria

UNIDA, NICARAGUA TRIUNFA!

Gobierno de Reconciliación y Unidad Nacional

El Pueblo, Presidente!

2017 TIEMPOS DE VICTORIAS!

Por Gracia de Dios!

Managua 19 de Septiembre del 2017

A QUIEN CONCIERNE

Por medio de la presente y a solicitud de la parte interesada, el compañero **Roger Deshon Meza**, Subdirector de la División de Informática de la Dirección General de Ingresos; certifica que:

La joven **Josseling J. Alemán Guido**, con cedula de identidad personal 001-050194-003C, desarrollo para esta institución desde 26 de Agosto 2016 hasta el 10 de Septiembre 2017, el proyecto denominado *"Implementación de Big Data en la Información tributaria de la Direccion General de Ingresos en el área de Fiscalización"* cuya tutoría fue asignada al Ing. Gabriel Lacayo docente de la Universidad Nacional de Ingeniería.

Debido a la naturaleza de los datos de los contribuyentes que ocupo la joven Alemán en el desarrollo del proyecto y a los protocolos de seguridad en cuanto a no poner en peligro la integridad, disponibilidad y confidencialidad de la información que maneja la institución, ella no podrá presentar el sistema en ejecución para el debido proceso de defensa de su trabajo monográfico. Permitiendo solamente hacer referencia al proyecto mediante imágenes y gráficos, ya sean del sistema o frutos del mismo.

Atentamente,

Dirección General de Ingresos – REPUBLICA DE NICARAGUA – AMERICA CENTRAL – División de Informática y Sistemas

Ing. Roger Deshon M.

Subdirector de la División de Informática

Dirección General de Ingresos

TIEMPOS DE VICTORIAS! Por Gracia de Dios!

FE, FAMILIA Y COMUNIDAD!

CRISTIANA, SOCIALISTA, SOLIDARIA!

DIRECCIÓN GENERAL DE INGRESOS

Costado Norte de Catedral Metropolitana - 2248-9999

www.dgi.gob.ni